ΤΕΧΝΟΛΟΓΙΚΟ ΕΚΠΑΙΔΕΥΤΙΚΟ ΙΔΡΥΜΑ ΚΡΗΤΗΣ

ΣΧΟΛΗ ΔΙΟΙΚΗΣΗΣ ΚΑΙ ΟΙΚΟΝΟΜΙΑΣ

ΤΜΗΜΑ ΛΟΓΙΣΤΙΚΗΣ

ΕΞΟΡΥΞΗ ΓΝΩΣΗΣ ΚΑΙ ΔΕΔΟΜΕΝΩΝ. ΠΛΕΟΝΕΚΤΗΜΑΤΑ ΚΑΙ ΜΕΙΟΝΕΚΤΗΜΑΤΑ ΣΕ ΜΙΑ ΕΠΙΧΕΙΡΗΣΗ.

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Φοιτήτρια: Ελένη Παρασύρη, Α.Μ. 6473

Επιβλέπων καθηγητής: Αντώνης Ποτηράκης

ΗΡΑΚΛΕΙΟ
2014

TECHNOLOGICAL EDUCATION INSTITUTE OF CRETE

SCHOOL OF MANAGEMENT AND ECONOMICS

DEPARTMENT OF LOGISTICS

KNOWLEDGE AND DATA EXTRACTION. ADVANTAGES AND DISADVANTAGES IN BUSINESS.

DIPLOMA THESIS

Student: Eleni Parasiri, A.M. 6473

Supervisor: Antonis Potirakis

HERAKLION
2014

Υπεύθυνη Δήλωση : *Βεβαιώνω ότι είμαι η συγγραφέας αυτής της πτυχιακής εργασίας και ότι κάθε βοήθεια την οποία είχα για την προετοιμασία της, είναι πλήρως αναγνωρισμένη και αναφέρεται στην πτυχιακή εργασία. Επίσης έχω αναφέρει τις όποιες πηγές από τις οποίες έκανα χρήση δεδομένων, ιδεών ή λέξεων, είτε αυτές αναφέρονται ακριβώς είτε παραφρασμένες. Επίσης βεβαιώνω ότι αυτή η πτυχιακή εργασία προετοιμάστηκε από εμένα προσωπικά ειδικά για τις απαιτήσεις του προγράμματος σπουδών του Τμήματος Λογιστικής του Τ.Ε.Ι. Κρήτης.*

ΠΕΡΙΛΗΨΗ

Στο σύγχρονο κόσμο των επιχειρήσεων και των πληροφοριακών συστημάτων, ο πλέον πολύτιμος πόρος για κάθε επιχείρηση, πέρα από το ανθρώπινο δυναμικό, είναι τα δεδομένα και η γνώση που μπορεί να εξαχθεί από αυτά. Η εποχή κατά την οποία κάθε επιχείρηση διέθετε χρόνο και χρήμα με σκοπό την βελτιστοποίηση των εσωτερικών της διαδικασιών έχει παρέλθει. Πρωταρχικός στόχος από τη σύλληψη ακόμη της ίδρυσης μιας εταιρείας είναι το κέρδος.

Πως λοιπόν τροφοδοτείται ένα σύστημα με τις κατάλληλες πληροφορίες οι οποίες θα βεβαιώσουν την ομαλή και επιτυχημένη πορεία μιας επιχείρησης; Πως συλλέγονται, αξιοποιούνται και μετατρέπονται σε χρήσιμες πληροφορίες τα δεδομένα που άπλετα υπάρχουν στον επιχειρηματικό κόσμο και πως αυτά τα δεδομένα μπορούν να κάνουν μια επιχείρηση ανταγωνιστική;

Εύκολα επομένως εξάγεται το συμπέρασμα ότι δύο σημαντικοί στόχοι για κάθε σύγχρονη επιχείρηση είναι η συλλογή δεδομένων και η εκμετάλλευση αυτών. Ο πρώτος από τους δύο επιτυγχάνεται σχετικά εύκολα μέσω εμπορικών εφαρμογών που έχουν αναπτυχθεί ειδικά για αυτό το σκοπό, ενώ ο δεύτερος αποτελεί αντικείμενο μελέτης για την επιστημονική περιοχή της εξόρυξης γνώσης από δεδομένα.

Την απάντηση στα παραπάνω ερωτήματα καλείται λοιπόν να δώσει η επιστήμη της εξόρυξης δεδομένων μέσω των ισχυρότατων εργαλείων και μεθόδων που διαθέτει. Η παρούσα εργασία έχει ως αντικείμενο να παρουσιάσει, να ερευνήσει και να αναλύσει τις μεθόδους και τις εφαρμογές της εξόρυξης δεδομένων, καθώς επίσης και όλα εκείνα τα πλεονεκτήματα και μειονεκτήματα που απορρέουν από αυτές.

ABSTRACT

In the modern world of business and IT systems, the most valuable resource for any business, apart from human resources, is data and knowledge that can be extracted from them. The era in which each company had time and money in order to optimize its internal processes has expired. The primary objective of capturing even the founding of a company is profit.

How then fed a system with appropriate information which will assure a smooth and successful a business? How this data can be collected, utilized and converted into useful information that vast amounts of data exist in the business world and how this data can make a business competitive?

Easily therefore, can be concluded that two important objectives for every modern business is data collection and exploitation thereof. The first of the two is achieved relatively easily through commercial applications developed especially for this purpose, while the second is the subject of study for the scientific area of data mining.

The answer to these questions is therefore invited to be given by the science of data mining via their strong tools and methods available. This study aimed to present, to investigate and analyze the methods and applications of data mining, as well as all those advantages and disadvantages arising from them.

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

ΚΕΦΑΛΑΙΟ 1
ΓΕΝΙΚΕΣ ΑΡΧΕΣ ΕΞΟΡΥΞΗΣ ΓΝΩΣΗΣ ΚΑΙ ΔΕΔΟΜΕΝΩΝ

ΕΙΣΑΓΩΓΗ

Σήμερα, οι οργανισμοί και οι επιχειρήσεις συγκεντρώνουν μεγάλο όγκο δεδομένων από διαφορετικές πηγές σε καθημερινή βάση. Παρόλα αυτά, η μετατροπή της μεγάλης ποσότητας δεδομένων σε γνώση και η δυνατότητα εκμετάλλευσης των δεδομένων ώστε να αποκτήσει κάποιος καλύτερη οπτική των καταστάσεων, παραμένει μία πρόκληση για τους περισσότερους οργανισμούς και επιχειρήσεις.

Για να μπορέσει κάποιος να φτάσει στην επίλυση περίπλοκων προβλημάτων, οι οποίες συνήθως, είναι καλά κρυμμένες στα ακατέργαστα δεδομένα και να αποκτήσει την κατάλληλη στρατηγική που θα πρέπει να ακολουθήσει ώστε να καταλήξει σε σωστές αποφάσεις, θα πρέπει να είναι σε θέση να εξάγει σημαντική γνώση ενώ ταυτόχρονα, να χρησιμοποιεί όλη τη διαθέσιμη πληροφορία που είναι αποθηκευμένη σε βάσεις δεδομένων και άλλες παρόμοιες πηγές.

Η εξόρυξη δεδομένων διευθετεί και επιλύει τα παραπάνω θέματα με την ανακάλυψη και αξιοποίηση προτύπων, δομών, μοντέλων, τάσεων και συσχετίσεων, με ένα αυτοματοποιημένο τρόπο. Η βέλτιστη απόκτηση γνώσης για μια επιχείρηση, έχει ως αποτέλεσμα το σχεδιασμό και τη λήψη των βέλτιστων αποφάσεων, την αύξηση της παραγωγικότητας σε στρατηγικά και επιχειρησιακά επίπεδα και κατά συνέπεια την αύξηση της ανταγωνιστικότητας της έναντι άλλων επιχειρήσεων.

1.1 ΟΡΙΣΜΟΣ ΕΞΟΡΥΞΗΣ ΓΝΩΣΗΣ ΚΑΙ ΔΕΔΟΜΕΝΩΝ

Η εξόρυξη γνώσης από δεδομένα (data mining) ή πιο απλά η εξόρυξη γνώσης είναι μια νέα δυναμική τεχνολογία που βοηθάει τις επιχειρήσεις να εστιάσουν στην σημαντική πληροφορία που βρίσκεται μέσα στις αποθήκες δεδομένων τους (data warehouses). Οι τεχνικές της είναι σε θέση να αναζητήσουν και να βρουν γρήγορα και λεπτομερειακά βάσεις δεδομένων για την αναζήτηση κρυμμένων προτύπων (patterns). Έτσι λοιπόν μπορούμε να πούμε ότι η εξόρυξη γνώσης είναι μια διαδικασία εξαγωγής κρυμμένης πληροφορίας από μεγάλες βάσεις δεδομένων.

«Εξόρυξη δεδομένων είναι η διαδικασία εξαγωγής υπονοούμενης και εν πολλοίς άγνωστης αλλά ενδεχομένως χρήσιμης γνώσης υπό την μορφή συσχετίσεων προτύπων και τάσεων, μέσω της εξέτασης ανάλυσης και επεξεργασίας βάσεων δεδομένων, συνδυάζοντας και χρησιμοποιώντας τεχνικές από την μηχανική μάθηση, την αναγνώριση προτύπων, την στατιστική, τις βάσεις δεδομένων και την οπτικοποίηση.» (Piatetsky-Shapiro & Frawley, 1991).

Παρά το γεγονός ότι υπάρχει μια γενικότερη συμφωνία ότι ο στόχος της εξόρυξης δεδομένων είναι η ανακάλυψη νέας και χρήσιμης πληροφορίας σε βάσεις δεδομένων, τα μέσα για την επίτευξη του στόχου αυτού ποικίλουν σε πολύ υψηλό βαθμό. Η εξόρυξη γνώσης περιλαμβάνει ένα ευρύ πεδίο υπολογιστικών μεθόδων που μεταξύ άλλων περιλαμβάνουν, την στατιστική ανάλυση (statistical analysis), τα δένδρα αποφάσεων (decision trees), τα νευρωνικά δίκτυα (neural networks), την εξαγωγή κανόνων (rule induction) και την γραφική οπτικοποίηση (graphic visualization).

Τέτοιες μέθοδοι χρησιμοποιούνται για την εύρεση συσχετίσεων, προτύπων και δομών σε μεγάλες και διαρκώς αυξανόμενες βάσεις δεδομένων. Ειδικά η εύρεση εργαλείων είναι ένα ιδιαίτερα σημαντικό εξαγόμενο της εξόρυξης δεδομένων μέσω σχέσεων μεταξύ των χαρακτηριστικών των βάσεων δεδομένων.

1.2 ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΑΝΕΥΡΕΣΗ ΓΝΩΣΗΣ

Η εξόρυξη γνώσης βοηθά τις σύγχρονες εταιρείες να εστιάζουν στα πιο σημαντικά στοιχεία από τις αποθήκες δεδομένων τους. Με άλλα λόγια είναι η διαδικασία εφαρμογής μεθόδων ανάλυσης σε μεγάλο όγκο δεδομένων. Ο χρήστης των εργαλείων εξόρυξης μπορεί να προβλέψει μελλοντικές συμπεριφορές και συνήθειες, ώστε οι εταιρίες να παίρνουν επιτυχημένες αποφάσεις. Συνειδητοποιούμε ότι οι τεχνικές εξόρυξης γνώσης αναπτύσσονται γρήγορα, δίχως αλλαγές στην υποδομή και με μοναδικό στόχο την αξιοποίηση των επεξεργασμένων δεδομένων.

Στη διεθνή βιβλιογραφία υπάρχει μια γενικότερη σύγχυση ανάμεσα στους όρους «Εξόρυξη Γνώσης» (Data mining) και «Ανεύρεση γνώσης στις βάσεις δεδομένων» (Knowledge discovery in databases, KDD). Σε πολλές περιπτώσεις αξίζει να σημειωθεί ότι οι δύο αυτοί όροι ταυτίζονται, ενώ στην πραγματικότητα η εξόρυξη δεδομένων αποτελεί τμήμα της ανεύρεσης γνώσης, συγκροτώντας το πυρήνα αυτής (Zaiane, 1999). Προκειμένου λοιπόν να κατανοηθεί καλύτερα η εξόρυξη δεδομένων, θα γίνει μια σύντομη αναφορά στη διαδικασία της ανεύρεσης γνώσης.

Η ανεύρεση γνώσης είναι μια επαναληπτική διαδικασία που αποτελείται από μια σειρά βημάτων, τα οποία οδηγούν από τη συλλογή των δεδομένων στην ανακάλυψη και εξαγωγή χρήσιμης πληροφορίας από αυτά.

Τα βήματα από τα οποία αποτελείται η διαδικασία ανεύρεσης γνώσης είναι τα ακόλουθα:

- **Καθαρισμός δεδομένων (Data cleaning):** Στο βήμα αυτό, αφαιρούνται από τη βάση δεδομένων αυτά που παράγουν θόρυβο, δηλαδή όλα εκείνα τα στοιχεία που μπορούν να επηρεάσουν ή και να διαστρεβλώσουν το αποτέλεσμα.

- **Ενσωμάτωση δεδομένων (Data integration):** Σε αυτό το βήμα τα δεδομένα που έχουν συλλεχθεί, πολλές φορές ανομοιογενή και από πολλές διαφορετικές πηγές, ενσωματώνονται σε μια κοινή βάση δεδομένων.

- **Επιλογή δεδομένων (Data selection):** Από όλα εκείνα τα δεδομένα που έχουμε στη διάθεση μας, επιλέγονται προσεκτικά εκείνα που είναι σχετικά και χρήσιμα για την ανάλυση που θα ακολουθήσει.

- **Τροποποίηση δεδομένων (Data transformation):** Τα δεδομένα που έχουμε επιλέξει δέχονται τις απαραίτητες τροποποιήσεις έτσι ώστε η μορφή τους να είναι κατάλληλη για την διαδικασία της εξόρυξης.

- **Εξόρυξη δεδομένων (Data mining):** Είναι το σημαντικότερο από τα βήματα της διαδικασίας και αυτό γιατί στο συγκεκριμένο στάδιο, ποικίλες εξελιγμένες τεχνικές χρησιμοποιούνται για την εξαγωγή δυνητικά χρήσιμων προτύπων.

- **Αξιολόγηση προτύπων (Pattern evaluation):** Στο βήμα αυτό αναγνωρίζονται χρήσιμα πρότυπα που αναπαριστούν γνώση, βάσει συγκεκριμένων μέτρων αξιολόγησης (evaluation measures).

- **Αναπαράσταση γνώσης (Knowledge representation):** Στο τελικό αυτό στάδιο, η γνώση που έχει ανακαλυφθεί παρουσιάζεται στον χρήστη, βοηθώντας τον έτσι να κατανοήσει και να ερμηνεύσει τα αποτελέσματα της εξόρυξης δεδομένων.

Πολλές φορές κάποια από τα παραπάνω βήματα μπορούν να συνδυαστούν μεταξύ τους για το καλύτερο δυνατό αποτέλεσμα. Για παράδειγμα, τα βήματα του καθαρισμού και της ενσωμάτωσης των δεδομένων, μπορούν να υλοποιηθούν μαζί με στόχο την δημιουργία μια αποθήκης δεδομένων. Με την ίδια λογική μπορούν να συνδυαστούν και τα βήματα της επιλογής και τροποποίησης των δεδομένων.

Από τα παραπάνω λοιπόν συμπεραίνουμε ότι η εξόρυξη δεδομένων είναι μια διαδικασία-κλειδί για την ανεύρεση γνώσης. Παρόλα αυτά, δεν καταλαμβάνει παρά μόνο ένα μικρό μέρος της όλης προσπάθειας, δεδομένου της πολυπλοκότητας της. Σε αυτό το σημείο αξίζει να σημειωθεί ότι ο χρήστης, εκμεταλλευόμενος την επαναληπτική μορφή της διαδικασίας ανεύρεσης γνώσης, έχει την δυνατότητα να τροποποιήσει τα μέτρα αξιολόγησης, να τελειοποιήσει την διαδικασία της εξόρυξης, να επιλέξει νέα δεδομένα, να τροποποιήσει περαιτέρω τα ήδη υπάρχοντα ή να ενσωματώσει στη βάση νέα από καινούργιες πηγές, με τελικό στόχο την εξαγωγή διαφορετικών και ακόμη πιο κατάλληλων αποτελεσμάτων.

1.3 ΣΤΟΧΟΙ ΤΗΣ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ

Οι μέθοδοι εξόρυξης γνώσης στοχεύουν στην ανακάλυψη στοιχειών που θα είναι χρήσιμα για τους οργανισμούς και τις επιχειρήσεις. Πληροφορίες για τυποποιημένες μορφές όπως για παράδειγμα, ότι υπάρχουν πελάτες που θα ψωνίσουν περισσότερο από δύο φορές σε περίοδο εκπτώσεων ή προσφορών, ή είναι πιθανό να αγοράσουν τουλάχιστον μια φορά κατά την διάρκεια των εορταστικών ημερών, Πάσχα και Χριστουγέννων, είτε για συσχετίσεις όπως όταν ένας πελάτης αγοράζει dvd player τότε πιθανότατα να αγοράσει και κάποια άλλη ηλεκτρονική συσκευή, μπορεί να αποτελέσουν καθοριστικούς παράγοντες για την λήψη αποφάσεων όσον αφορά τη λειτουργία μιας εμπορικής επιχείρησης. Αυτό συμβαίνει επειδή μπορεί να ληφθούν αποφάσεις σχετικά με το ωράριο, το ύψος και τη διάρκεια των εκπτώσεων, ακόμη και για την τοποθέτηση των προϊόντων μέσα στα καταστήματα.

Παράλληλα τέτοιου είδους πληροφορίες χρησιμοποιούνται για τον προγραμματισμό χρήσης πρόσθετων αποθηκευτικών χώρων ή και για τον σχεδιασμό διαφορετικών στρατηγικών μάρκετινγκ. Τα στελέχη της επιχείρησης, που είναι υπεύθυνα για την λήψη των αποφάσεων εκμεταλλεύονται τις δυνατότητες της εξόρυξης γνώσης και μετατρέπουν τις γνώσεις σε επιτυχή αποτελέσματα. Παρακάτω περιγράφονται και αναλύονται οι στόχοι της εξόρυξης δεδομένων.

Η εξόρυξη δεδομένων έχει λοιπόν σαν βασικούς της στόχους την εφαρμογή τεχνικών πρόβλεψης και συμπεριφοράς τάσεων (prediction), την αναγνώριση, την περιγραφή (description) σε μεγάλες βάσεις δεδομένων (Fayyad et al, 1996,1996, Hegland, 2003), καθώς επίσης την ταξινόμηση και την βελτιστοποίηση των πόρων της. Ειδικότερα:

- ✓ **Πρόβλεψη:** Περιλαμβάνει την χρήση μερικών μεταβλητών ή χαρακτηριστικών μιας βάσης δεδομένων για την πρόβλεψη άγνωστων ή μελλοντικών τιμών χρήσιμων μεταβλητών. Με άλλα λόγια, οι διαδικασίες πρόβλεψης της εξόρυξης δεδομένων (predictive data mining tasks), προσπαθούν να κάνουν εκτιμήσεις βγάζοντας συμπεράσματα από τα διαθέσιμα δεδομένα. Η προσπάθεια πρόβλεψης μελλοντικών συμπεριφορών έχει ως στόχο να ληφθούν αποφάσεις που να μεγιστοποιούν το κέρδος και να προλαμβάνουν δυσάρεστες καταστάσεις. Τα αποτελέσματα της εξόρυξης μπορεί να είναι πληροφορίες σχετικές με το ύψος των πωλήσεων ενός καταστήματος για μια συγκεκριμένη χρονική περίοδο, αλλά και αν το κλείσιμο μιας γραμμής παραγωγής θα είχε θετική επίδραση στις πωλήσεις. Συγχρόνως σε επιστημονικό επίπεδο, η μελέτη παλαιότερων σεισμικών φαινόμενων ίσως να οδηγούσε στην πρόβλεψη σεισμικής δραστηριότητας.

- ✓ **Αναγνώριση:** Σε αυτή τη φάση οι τυποποιημένες μορφές των δεδομένων χρησιμοποιούνται για να δείξουν την ύπαρξη μιας δραστηριότητας ή ενός γεγονότος.

- ✓ **Περιγραφή:** Είναι η διαδικασία η οποία επικεντρώνεται στην ανακάλυψη προτύπων και αναπαριστά τα δεδομένα μιας πολύπλοκης βάσης δεδομένων με όσο το δυνατό πιο κατανοητό και αξιοποιήσιμο τρόπο. Με άλλα λόγια, οι περιγραφικές διαδικασίες της εξόρυξης δεδομένων (descriptive data mining tasks) περιγράφουν τις γενικές ιδιότητες των υπαρχόντων διαθέσιμων δεδομένων.

- ✓ **Ταξινόμηση:** Σε αυτό το στάδιο έχουμε διαχωρισμό των στοιχείων, με αποτέλεσμα να προκύπτουν διαφορετικές κατηγορίες ή κλάσεις. Για παράδειγμα, οι πελάτες ενός σούπερ μάρκετ είναι δυνατόν να χωριστούν σε παρορμητικούς, πιστούς ή αλλιώς όπως θα λέγαμε κανονικούς, σπάνιους και σε φίλους των εκπτώσεων και προσφορών. Κατά την ανάλυση των πωλήσεων αυτή η κατηγοριοποίηση χρησιμοποιείται για να ληφθούν αποφάσεις, ώστε να προσελκυστούν περισσότεροι πελάτες ανεξαρτήτως κατηγορίας.

- ✓ **Βελτιστοποίηση:** Μεταξύ των άλλων σκοπός της εξόρυξης γνώσης είναι η βέλτιστη χρήση κάποιων πόρων κάτω από περιορισμούς. Τέτοιοι πόροι μπορεί να είναι ο χρόνος, ο χώρος, το χρήμα και η μεγιστοποίηση κάποιων μεγεθών, όπως είναι τα κέρδη είτε οι πωλήσεις. Σε αυτή την περίπτωση η εξόρυξη γνώσης έχει κοινά σημεία με την επιχειρησιακή έρευνα.

1.4 ΔΙΑΔΙΚΑΣΙΑ ΕΞΟΡΥΞΗΣ ΓΝΩΣΗΣ

Η διαδικασία ανακάλυψης γνώσης από βάσεις δεδομένων (KDD) συνήθως ορίζεται από τα εξής στάδια:

1. Συλλογή
2. Προ επεξεργασία
3. Μετασχηματισμός
4. Εξόρυξη δεδομένων
5. Ερμηνεία και Αξιολόγηση

Υπάρχουν και παραλλαγές για τον ορισμό των σταδίων αυτών σύμφωνα και με το Cross Industry Standard Process for Data Mining (CRISP-DM) όπου τα στάδια έχουν ως εξής:

1. Κατανόηση Θέματος
2. Κατανόηση δεδομένων
3. Προετοιμασία δεδομένων
4. Μοντελοποίηση
5. Αξιολόγηση
6. Ανάπτυξη

Στην πιο απλοποιημένη διαδικασία της διαχωρίζεται στα εξής στάδια:

1. Προ-επεξεργασία
2. Εξόρυξη δεδομένων
3. Επικύρωση αποτελέσματος.

Σε αυτό το σημείο αξίζει να κάνουμε μια αναφορά σε δύο βασικά στάδια που περιλαμβάνει η διαδικασία εξόρυξης γνώσης, αυτό της προ-επεξεργασίας και της μοντελοποίησης.

1.4.1 ΠΡΟ-ΕΠΕΞΕΡΓΑΣΙΑ

Πριν την εφαρμογή των αλγορίθμων εξόρυξης δεδομένων, το ερευνώμενο σύνολο αυτών πρέπει να συναρμολογείται. Καθώς η εξόρυξη δεδομένων μπορεί να αποκαλύψει μόνο τα πρότυπα που πράγματι εμφανίζονται στα δεδομένα, το φάσμα αυτών που ερευνούμε, πρέπει να είναι αρκετά ευρύ για να περιέχει αυτά τα πρότυπα προκειμένου να προκύψει σε ένα αποδεκτό χρονικό διάστημα. Η προ επεξεργασία είναι απαραίτητη για την ανάλυση πολλών παραγόντων-συνόλων πριν την εξόρυξη δεδομένων. Έτσι το ερευνώμενο σύνολο καθαρίζεται. Το καθάρισμα δεδομένων διαγράφει τις παρατηρήσεις που περιέχουν θόρυβο και αυτές με ελλιπή δεδομένα.

Η εξόρυξη δεδομένων περιλαμβάνει έξι κατηγορίες:

- **Ανίχνευση ανωμαλιών (Anomaly detection):** Ο προσδιορισμός ασυνήθιστων εγγραφών δεδομένων, που μπορεί να παρουσιάζουν κάποιο ενδιαφέρον ή λάθη στα δεδομένα που απαιτούν περαιτέρω έρευνα.

- **Κατηγοριοποίηση:** Είναι η διαδικασία γενίκευσης γνωστών δομών για την εφαρμογή της πάνω σε νέα δεδομένα. Για παράδειγμα, ένα πρόγραμμα ηλεκτρονικού ταχυδρομείου ενδέχεται να προσπαθήσει να χαρακτηρίσει ένα μήνυμα ηλεκτρονικού ταχυδρομείου ως νόμιμο ή spam.

- **Συσταδιοποίηση:** Πρόκειται για τη διαδικασία ανακάλυψης ομάδων και δομών στα δεδομένα που είναι «παρόμοια» κατά κάποιο τρόπο, χωρίς να χρησιμοποιούνται γνωστές δομές στα δεδομένα.

- **Ανάλυση συσχέτισης (Μοντέλο αλληλεξάρτησης):** Αναζητήσεις για σχέσεις μεταξύ των μεταβλητών. Για παράδειγμα, ένα σούπερ μάρκετ μπορεί να συλλέξει δεδομένα που αφορούν της αγοραστικές συνήθειες των πελατών του. Χρησιμοποιώντας τους κανόνες συσχέτισης, μπορεί να υπολογίσει ποια προϊόντα αγοράζονται συνήθως μαζί και να χρησιμοποιήσει αυτή την πληροφορία για αγοραστικούς σκοπούς προς όφελος των πελατών του και του ίδιου.

- **Παλινδρόμηση:** Προσπαθεί να βρει μία συνάρτηση που μοντελοποιεί τα δεδομένα με το λιγότερο δυνατό λάθος.

- **Σύνοψη:** Παρέχει μια συμπαγέστερη αναπαράσταση των δεδομένων, συμπεριλαμβάνοντας την οπτικοποίηση και την παραγωγή κανόνων.

1.4.2 ΜΟΝΤΕΛΟΠΟΙΗΣΗ

Η τεχνική που εφαρμόζεται για να μάθουμε από την εξόρυξη γνώσης πληροφορίες που δεν γνωρίζουμε ή που θα συμβούν στο μέλλον ονομάζεται μοντελοποίηση. Δηλαδή η κατασκευή ενός μοντέλου για μια κατάσταση που γνωρίζουμε την απάντηση και στη συνέχεια η εφαρμογή του σε μια άλλη που δεν ξέρουμε.

Για παράδειγμα, αν αναζητούσαμε μια βυθισμένη ισπανική γαλέρα στην ανοικτή θάλασσα το πρώτο πράγμα που ίσως σκεφτόμασταν θα ήταν να ερευνήσουμε όλες τις περασμένες περιπτώσεις εύρεσης ισπανικών θησαυρών από άλλους. Ίσως λοιπόν να παρατηρούσαμε ότι αυτά τα πλοία στην πλειονότητα τους βρέθηκαν στις ακτές Βερμούδα και ότι υπήρχαν κάποιες βέβαιες πορείες που ακολουθούσαν οι καπετάνιοι των πλοίων αυτών εκείνη την εποχή. Αυτές οι ομοιότητες σημειώνονται και κτίζεται ένα μοντέλο που περιλαμβάνει τα χαρακτηριστικά που είναι κοινά στις τοποθεσίες αυτών των βυθισμένων θησαυρών. Με αυτό το μοντέλο αρχίζει το ψάξιμο σε περιοχές που δείχνει αυτό ότι είναι πιθανό να υπήρξε μια παρόμοια κατάσταση στο παρελθόν. Αν το μοντέλο είναι καλό ο θησαυρός θα βρεθεί. Η σκέψη κτισίματος μοντέλων από τους ανθρώπους υπήρχε αρκετό καιρό πριν από την τεχνολογία της εξόρυξης γνώσης. Η διαδικασία που ακολουθείται είναι να φορτώνονται οι υπολογιστές με στοιχεία για πολλές καταστάσεις ενώ μια απάντηση είναι γνωστή. Έπειτα το λογισμικό εξόρυξης γνώσης τρέχει πάνω σε αυτό τα δεδομένα και διαλέγει τα πιο χαρακτηριστικά που θα συμπεριληφθούν στο μοντέλο. Όταν τελειώσει η κατασκευή του μοντέλου είναι δυνατό να χρησιμοποιηθεί σε παρόμοιες καταστάσεις που δεν γνωρίζουμε την απάντηση.

Για παράδειγμα ας υποθέσουμε ότι βρισκόμαστε στη θέση του διευθυντή μάρκετινγκ μιας εταιρίας τηλεπικοινωνιών και θέλουμε να αποκτήσουμε μερικούς πελάτες που κάνουν τηλεφωνήματα μεγάλων αποστάσεων. Βρισκόμαστε δηλαδή αντιμέτωποι με ένα πρόβλημα απόφασης, σε ποιους να απευθυνθούμε. Θα μπορούσαμε να ταχυδρομήσουμε με τυχαίο τρόπο κουπόνια στο γενικό πληθυσμό όπως θα μπορούσαμε να ταξιδεύουμε στις θάλασσες ψάχνοντας για βυθισμένους θησαυρούς. Πάντως σε καμιά από τις δυο περιπτώσεις δεν θα είχαμε τα επιθυμητά αποτελέσματα. Αντί αυτού, θα μπορούσαμε να χρησιμοποιήσουμε την εμπειρία της εταιρίας που βρίσκεται αποθηκευμένη στις βάσεις δεδομένων και να κτίσουμε ένα μοντέλο. Ο διευθυντής μάρκετινγκ έχει πρόσβαση σε πολλές πληροφορίες σχετικές με τους πελάτες όπως η ηλικία τους, το φύλο, αν είναι καλοί πληρωτές, το πόσα τηλεφωνήματα μεγάλων αποστάσεων κάνουν. Το πρόβλημα είναι ότι δεν γνωρίζουμε πόσο πολύ θα κάνουν χρήση τηλεφωνημάτων σε απομακρυσμένες περιοχές. Επειδή θέλουμε αυτούς που κάνουν πολλά από αυτά τα τηλεφωνήματα, μπορούμε να το πετύχουμε αυτό κτίζοντας ένα μοντέλο. Ένα τέτοιο απλό μοντέλο που θα ταίριαζε σε μια τηλεπικοινωνιακή εταιρία είναι το παρακάτω:

Για παράδειγμα το 98% των πελατών που έχουν λογαριασμό μεγαλύτερο από 6000 ευρώ το χρόνο δαπανούν περισσότερα από 80 ευρώ το μήνα για τηλεφωνήματα σε μακρινές περιοχές. Αυτό το μοντέλο θα μπορούσε να εφαρμοστεί στα δεδομένα των πιθανών πελατών και να δοθεί απάντηση στο πρόβλημα απόφασης. Αφού γίνει αυτό θα ξέρει σε ποιους θα πρέπει να απευθυνθεί η εταιρεία.

Η εξόρυξη γνώσης με άλλα λόγια είναι μια επέκταση της στατιστικής με κάποια στοιχεία τεχνητής νοημοσύνης και μηχανικής μάθησης(machine learning). Η εξόρυξη γνώσης είναι μια τεχνολογία και όπως και η στατιστική δεν αποτελεί επιχειρηματική λύση. Είναι μόνο μια τεχνολογία. Για παράδειγμα, σε περίπτωση που έχουμε ένα κατάλογο εμπόρων λιανικής πώλησης και πρέπει να αποφασιστεί ποιοι από αυτούς θα ενημερωθούν για ένα νέο προϊόν. Η εξόρυξη γνώσης αναζητά την πληροφορία που βρίσκεται μέσα στις βάσεις δεδομένων προηγούμενων συναλλαγών με τους πελάτες καθώς και σε χαρακτηριστικά αυτών, όπως αν ανταποκρίθηκαν στο παρελθόν, η ηλικία τους, η διεύθυνση τους κλπ. Το λογισμικό της εξόρυξης γνώσης χρησιμοποιεί αυτά τα στοιχεία για να κατασκευάσει ένα μοντέλο συμπεριφοράς του πελάτη έτσι ώστε αυτό να χρησιμοποιηθεί για να προβλεφθεί ποιοι πελάτες θα ανταποκριθούν στο νέο προϊόν. Επομένως ένα στέλεχος του τμήματος marketing μπορεί να επιλέξει τους πιθανούς πελάτες. Αντιλαμβανόμαστε ότι το λογισμικό της εταιρείας έχει την δυνατότητα να τροφοδοτεί τα κατάλληλα σημεία επαφής (web servers, τηλεφωνικά κέντρα, e-mails κλπ) με τις αποφάσεις έτσι ώστε οι πελάτες να παίρνουν τις πληροφορίες που χρειάζονται.

Παρακάτω βλέπουμε τα στάδια που μεσολαβούν μέχρι να είναι δυνατή η ερμηνεία και η ανάλυση των αποτελεσμάτων. Άρα η ανακάλυψη γνώσης, ή αλλιώς η διαδικασία καθορισμού και επίτευξης ενός σκοπού μέσω επαναληπτικής εξόρυξης γνώσης, αποτελείται από τα εξής τρία στάδια:

- Προετοιμασία των δεδομένων
- Υλοποίηση και αποτίμηση του μοντέλου
- Ανάπτυξη του μοντέλου

Αρχικά ο αναλυτής προετοιμάζει ένα σύνολο στοιχείων για να κτιστεί ένα σωστό μοντέλο στις επόμενες φάσεις. Στοχεύοντας τις αναγκαίες πληροφορίες για μια επιχείρηση, ένα σωστό μοντέλο θα προβλέπει τι πιθανότητα υπάρχει ο πελάτης να αγοράσει προϊόντα από έναν νέο κατάλογο. Οι προβλέψεις βασίζονται σε παράγοντες που επιδρούν τις αγορές των πελατών και γι' αυτό ένα μοντέλο συνόλου δεδομένων θα έπρεπε να περιέχει όλους τους πελάτες που ανταποκρίθηκαν σε καταλόγους μέσω ταχυδρομείων, e-mails κλπ. τα τελευταία 4 χρόνια, τα 8 πιο ακριβά προϊόντα που αγόρασε κάθε πελάτης, τις δημογραφικές πληροφορίες τους, και στοιχεία για τους καταλόγους που έγιναν οι αγορές. Συνειδητοποιούμε ότι πολύπλοκες ερωτήσεις με μεγάλες απαντήσεις περιλαμβάνονται στην προετοιμασία των δεδομένων.

Για παράδειγμα για την εταιρία που αναφερθήκαμε προηγουμένως, η προετοιμασία του μοντέλου έχει συνδέσεις μεταξύ του πίνακα πωλήσεων και του πίνακα πελατών, καθώς και για τον προσδιορισμό των 8 κορυφαίων προϊόντων για κάθε πελάτη. Επομένως η αποτελεσματική επεξεργασία ερωτήσεων υποστήριξης αποφάσεων σχετίζονται με το περιβάλλον της εξόρυξης γνώσης. Η εξόρυξη γνώσης περιλαμβάνει την επαναληπτική κατασκευή μοντέλων πάνω σε ένα σύνολο δεδομένων που έχει προετοιμαστεί και εν συνεχεία στην ανάπτυξη ενός ή περισσοτέρων μοντέλων. Εκτός των άλλων οι αναλυτές - ειδικοί εργάζονται με επαναληπτικό τρόπο σε δείγματα συνόλων δεδομένων, επειδή το κτίσιμο των μοντέλων σε μεγάλα σύνολα δεδομένων είναι αρκετά ακριβό. Ο αναλυτής κατασκευάζει το μοντέλο πάνω στο σύνολο δεδομένων, αφού όμως πρώτα έχει αποφασιστεί ποιο μοντέλο θα χρησιμοποιηθεί.

Στη φάση της υλοποίησης εντοπίζονται οι τυποποιημένες μορφές που ορίζουν ένα χαρακτηριστικό-στόχος (target attribute). Αν και μερικές κλάσεις μοντέλων εξόρυξης γνώσης συμβάλλουν σημαντικά στην πρόβλεψη τόσο κρυφών χαρακτηριστικών όσο και φανερά καθορισμένων, κρίνονται αναγκαία, για την επιλογή του μοντέλου τα χαρακτηριστικά της ακρίβειας και της αποτελεσματικότητας του αλγορίθμου κατασκευής του μοντέλου σε μεγάλα σύνολα δεδομένων. Αξιοπρόσεκτο είναι το γεγονός ότι από στατιστικής πλευράς η ακρίβεια των περισσότερων μοντέλων βελτιώνεται με το πλήθος των δεδομένων που χρησιμοποιούνται.

1.5 ΜΕΘΟΔΟΙ ΕΞΟΡΥΞΗΣ ΓΝΩΣΗΣ ΚΑΙ ΔΕΔΟΜΕΝΩΝ

Οι βασικότερες από τις μεθόδους της εξόρυξης δεδομένων (Fayyad et al, 1996, Berry & Linoff, 2004), μέσω των οποίων επιτυγχάνονται οι στόχοι που αναφέραμε προηγουμένως, είναι οι εξής:

- Κατηγοριοποίηση
- Συσταδιοποίηση
- Ανάλυση Συσχέτισης
- Παλινδρόμηση

1.5.1 ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ

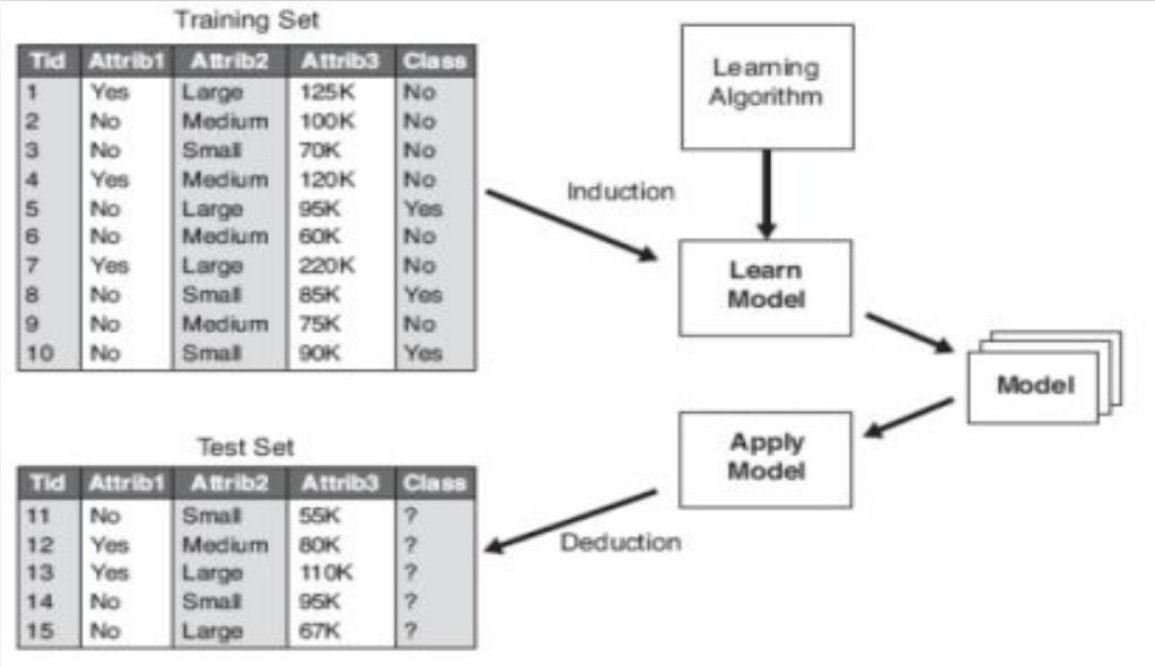
Η διαδικασία της κατηγοριοποίησης, ή αλλιώς ταξινόμησης (classification) περιλαμβάνει την οργάνωση ενός συνόλου αντικειμένων (objects) που περιγράφονται από ένα σύνολο χαρακτηριστικών (attributes), σε μια σειρά από προκαθορισμένες κλάσεις (classes), χρησιμοποιώντας μεθόδους μάθησης με επίβλεψη (supervised learning methods). Οι τεχνικές της ταξινόμησης ή αλλιώς κατηγοριοποίησης χρησιμοποιούν κατά κανόνα ένα σύνολο εκπαίδευσης (training set), όπου όλα τα αντικείμενα είναι ήδη συνδεδεμένα με γνωστές κλάσεις. Ο αλγόριθμος ταξινόμησης μαθαίνει από αυτό το σύνολο, χρησιμοποιώντας την μάθηση αυτή για την κατασκευή ενός μοντέλου και το μοντέλο αυτό στην συνέχεια ταξινομεί νέα αντικείμενα στις κατάλληλες κλάσεις (Fayyad et al, 1996, Zaiane, 1999, Kotsiantis, 2007). Άρα μπορούμε να πούμε ότι η κατηγοριοποίηση μαθαίνει σε μία λειτουργία να χαρτογραφεί ή πιο απλά να ταξινομεί ένα στοιχείο δεδομένων σε μία από τις διάφορες προκαθορισμένες κατηγορίες. Η κατηγοριοποίηση πρόκειται ίσως για την πιο δημοφιλή τεχνική με πλήθος εφαρμογών στην αναγνώριση προτύπων και εικόνας σε διάφορους κλάδους.

Στην πράξη μια διαδικασία κατηγοριοποίησης μπορεί να οριστεί ως η εκτέλεση δύο συγκεκριμένων βημάτων:

1. Δημιουργία μοντέλου βασιζόμενου σε δεδομένα εκπαίδευσης
2. Εφαρμογή του μοντέλου στο σύνολο των δεδομένων

Αν και βάσει του επιστημονικού ορισμού το δεύτερο από τα παραπάνω βήματα είναι αυτό της κατηγοριοποίησης, το πρώτο είναι το βήμα που απαιτεί και την μεγαλύτερη προσπάθεια. Η εργασία της κατηγοριοποίησης χαρακτηρίζεται από έναν καλά καθορισμένο ορισμό των κατηγοριών και το σύνολο που χρησιμοποιείται για την εκπαίδευση του μοντέλου αποτελείται από προ κατηγοριοποιημένα παραδείγματα.

Η επόμενη εικόνα δείχνει μία γενική προσέγγιση επίλυσης προβλήματος κατηγοριοποίησης. Αρχικά, πρέπει να δοθεί ένα training set το οποίο περιέχει εγγραφές των οποίων οι ετικέτες κατηγορίας είναι σωστές. Το training set χρησιμοποιείται για να φτιάξει το μοντέλο ταξινόμησης, το οποίο μετέπειτα εφαρμόζεται στο test set, όπου περιέχει εγγραφές των οποίων οι ετικέτες κατηγορίας είναι άγνωστες. Η διαδικασία που ακολουθείται δηλαδή έχει να κάνει με την παραγωγή της test set με άγνωστες ετικέτες από το αρχικό training set οι οποίες πρέπει να προβλεφθούν από κάποιον αλγόριθμο με όσο το δυνατό μεγαλύτερη επιτυχία. Σκοπός της παρούσας μελέτης είναι οι βελτιωμένες προβλέψεις των test set που περιλαμβάνουν δεδομένα επιχειρήσεων εισηγμένων στο ΧΑΑ.



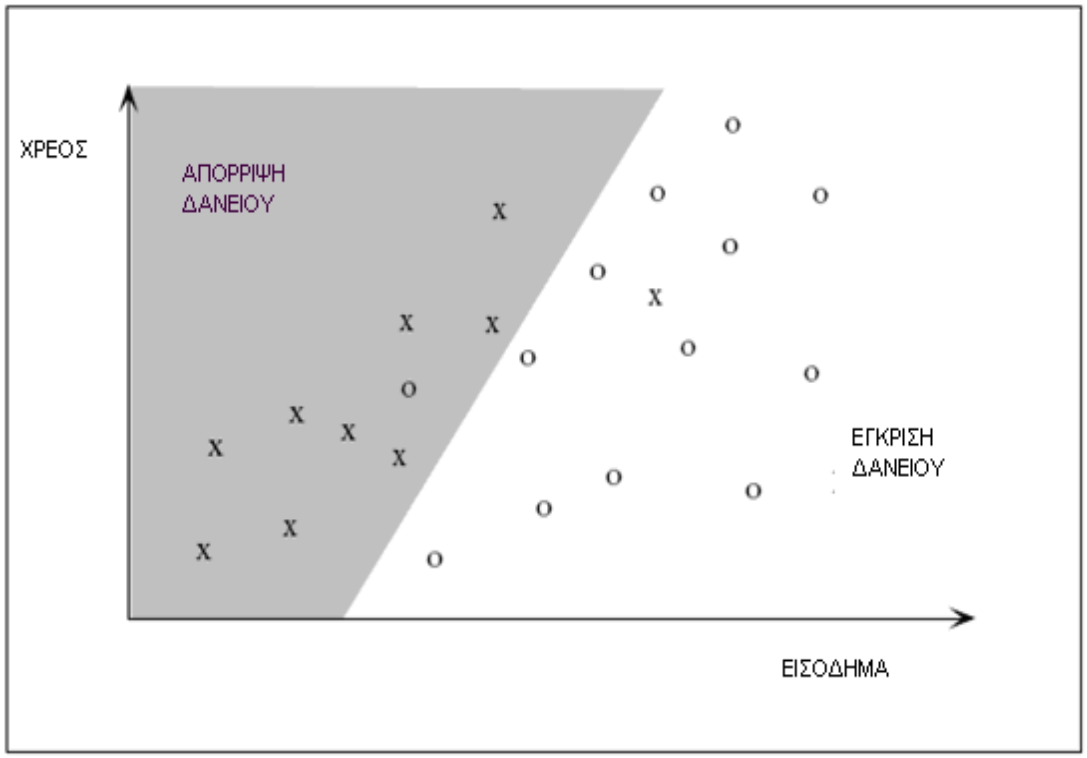
Η αξιολόγηση της απόδοσης ενός μοντέλου ταξινόμησης βασίζεται στον αριθμό των εγγραφών του test set που προβλέφθηκαν σωστά ή λάθος από τον ταξινομητή. Για να είναι ευκολότερη η σύγκριση των αποδόσεων διαφορετικών μοντέλων χρησιμοποιούνται δύο δείκτες επίδοσης, η ακρίβεια (accuracy) και η αποτίμηση του σφάλματος (error rate).

$$\text{Ακρίβεια} = \frac{\text{Σωστές προβλέψεις}}{\text{Σύνολο προβλέψεων}}$$

$$\text{Αποτίμηση σφάλματος} = \frac{\text{Λάθος προβλέψεις}}{\text{Σύνολο προβλέψεων}}$$

Έτσι τελικά ο ταξινομητής με τη μεγαλύτερη ακρίβεια και το μικρότερη αποτίμηση σφάλματος είναι ορθότερος και πιο αποτελεσματικός, δηλαδή μπορεί και κάνει καλύτερες προβλέψεις.

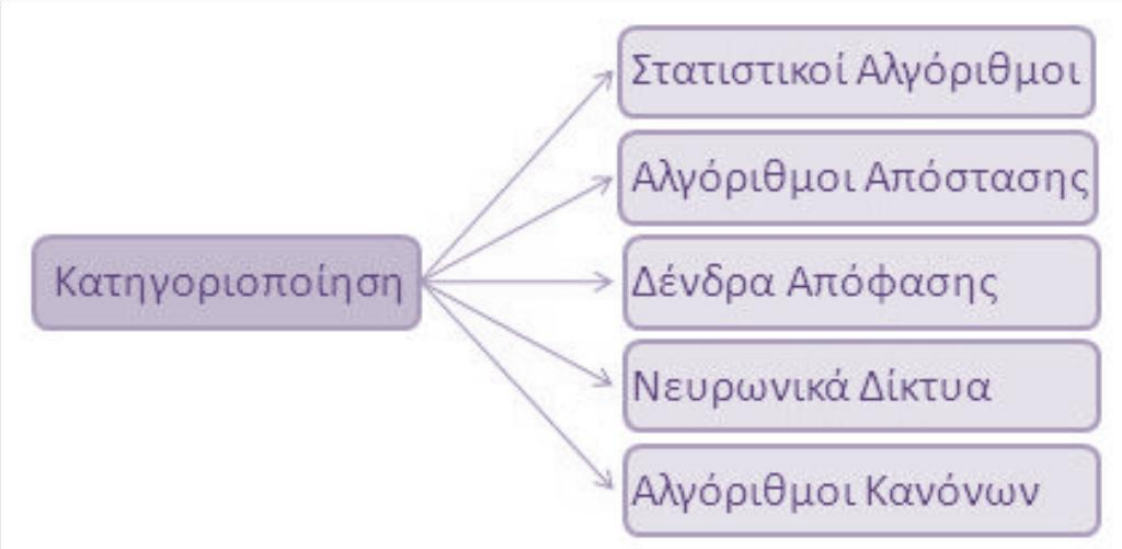
Στην παρακάτω εικόνα έχουμε έναν απλό διαχωρισμό των στοιχείων δανείου σε δύο περιοχές κατηγοριών. Η τράπεζα πιθανώς να θελήσει να χρησιμοποιήσει τις περιοχές ταξινόμησης για να αποφασίσει, εάν θα δοθεί δάνειο ή όχι στους μελλοντικούς υποψηφίους.



Ένα απλό γραμμικό όριο κατηγοριοποίησης για το σύνολο των στοιχείων δανείου. Η διαμορφωμένη περιοχή δείχνει την κατηγορία (απόρριψη-έγκριση) και όχι το δάνειο.

Η εργασία της κατηγοριοποίησης χαρακτηρίζεται από έναν καλά καθορισμένο ορισμό των κατηγοριών και το σύνολο που χρησιμοποιείται για την εκπαίδευση του μοντέλου αποτελείται από προ κατηγοριοποιημένα παραδείγματα. Η βασική εργασία είναι να δημιουργηθεί ένα μοντέλο το οποίο θα μπορούσε να εφαρμοστεί για να οργανώσει δεδομένα που δεν έχουν ακόμα κατηγοριοποιηθεί. Στις περισσότερες περιπτώσεις, υπάρχει ένα περιορισμένος αριθμός κατηγοριών και εμείς θα πρέπει να αναθέσουμε κάθε εγγραφή στην κατάλληλη κατηγορία. Για αυτό το σκοπό χρησιμοποιούνται κάποιες τεχνικές, τις οποίες μπορούμε να κατατάξουμε σε δύο βασικές κατηγορίες. Η πρώτη χρησιμοποιεί τα λεγόμενα δέντρα απόφασης (decision trees) ενώ η δεύτερη τα νευρωνικά δίκτυα (neural networks).

Γενικά μπορούμε να πούμε ότι οι αλγόριθμοι κατηγοριοποίησης μπορούν να διαχωριστούν στις ακόλουθες κατηγορίες, κάποιες από τις οποίες θα αναλύσουμε στη συνέχεια:



1.5.1.1 ΔΕΝΔΡΑ ΑΠΟΦΑΣΗΣ

Τα δέντρα απόφασης (decision trees) είναι μια από τις πιο σημαντικές και ευρύτατα διαδεδομένες μεθόδους για την ταξινόμηση δεδομένων. Σύμφωνα με τους Quinlan (1986, 1987, 1993) και Murphy (1998), τα δέντρα απόφασης είναι δομές που ταξινομούν τα αντικείμενα μιας βάσης δεδομένων βάσει των τιμών των χαρακτηριστικών αυτών. Η κατασκευή του από την άλλη βασίζεται σε ένα σύνολο εκπαίδευσης, το οποίο περιλαμβάνει προ-ταξινομημένα δεδομένα.

Η ταξινόμηση ενός νέου αντικειμένου μέσω ενός δέντρου απόφασης ακολουθεί κάποια βήματα. Ξεκινώντας από την ρίζα του δέντρου (αρχικός κόμβος) και εξετάζοντας τα χαρακτηριστικά που καθορίζονται από τον κόμβο αυτό, προσδιορίζονται διαδοχικά οι εσωτερικοί κόμβοι του δέντρου που πρέπει να ακολουθηθούν, έως ότου καταλήξουμε σε ένα συγκεκριμένο φύλλο. Πολλά διαφορετικά φύλλα μπορούν να οδηγούν στην ίδια ταξινόμηση, αλλά κάθε φύλλο κάνει την ταξινόμηση αυτή για διαφορετικό λόγο. Σε κάθε εσωτερικό κόμβο, εξετάζεται αν το προς ταξινόμηση αντικείμενο ικανοποιεί τον συγκεκριμένο κόμβο. Η έκβαση της εξέτασης αυτής καθορίζει το κλαδί που θα ακολουθηθεί στην συνέχεια, καθώς και τον επόμενο κόμβο. Η κλάση στην οποία θα ταξινομηθεί το νέο αντικείμενο αντιστοιχεί σε ένα από τα φύλλα του δέντρου απόφασης, είναι δε αυτή του τελικού κόμβου (Mitchell, 1997).

Μερικά από τα κρίσιμα ζητήματα που αφορούν τους αλγόριθμους δημιουργίας δένδρων απόφασης ή κατηγοριοποίησης είναι τα ακόλουθα:

- ✓ **Η επιλογή των γνωρισμάτων διάσπασης:** Τα γνωρίσματα του συνόλου δεδομένων που θα επιλεγούν για τη δημιουργία του δένδρου είναι κρίσιμης σημασίας. Αναμφισβήτητα, κάποια γνωρίσματα είναι σημαντικότερα από κάποια άλλα και η επιλογή των καταλληλότερων είναι πολλές φορές όχι μόνο θέμα εξέτασης των δεδομένων εκπαίδευσης αλλά και εμπεριστατωμένης άποψης ειδικών πάνω στη φύση των δεδομένων.

- ✓ **Η διάταξη των γνωρισμάτων διάσπασης:** Εκτός από το ποια γνωρίσματα είναι πλέον κατάλληλα, κρίσιμη απόφαση είναι και η διάταξη των καταλληλότερων. Μια λάθος διάταξη γνωρισμάτων διάσπασης μπορεί να σημαίνει τον επανέλεγχο γνωρισμάτων αρκετές φορές.

- ✓ **Οι διασπάσεις:** Γνωρίσματα με μικρό πεδίο τιμών οδηγούν σε φανερό αριθμό διασπάσεων, σε αντίθεση με γνωρίσματα συνεχών πεδίων όπου ο αριθμός των διασπάσεων κάθε άλλο παρά εύκολος μπορεί να θεωρηθεί.

- ✓ **Η δομή του δένδρου:** Η δομή του δένδρου απόφασης ασφαλώς και παίζει σημαντικό ρόλο στη δεύτερη από τις δύο φάσεις κατηγοριοποίησης, αυτή της εφαρμογής του δένδρου πάνω στις πλειάδες της βάσης δεδομένων. Ισοζυγισμένα δένδρα λίγων επιπέδων ασφαλώς και βοηθούν στην αποδοτικότερη κατηγοριοποίηση.

- ✓ **Τα κριτήρια του τερματισμού:** Η δημιουργία του δένδρου απόφασης ολοκληρώνεται όταν τα δεδομένα κατηγοριοποιούνται με απόλυτη ακρίβεια. Αυτό ωστόσο κρύβει κινδύνους στη δημιουργία μεγάλων δένδρων. Από αυτό συμπεραίνουμε ότι ο συμβιβασμός μεταξύ ακρίβειας της κατηγοριοποίησης και αποδοτικότητας του δένδρου είναι απαραίτητος.

- ✓ **Τα δεδομένα εκπαίδευσης:** Η δημιουργία του δένδρου βασίζεται αποκλειστικά στα δεδομένα εκπαίδευσης. Μικρό σύνολο τέτοιων δεδομένων ίσως οδηγήσει σε δένδρο μη κατάλληλο για το σύνολο των δεδομένων που διαθέτουμε. Μεγάλο σύνολο δεδομένων εκπαίδευσης μπορεί να προκαλέσει υπερ-προσπαρμογή.

- ✓ **Το κλάδεμα του δένδρου:** Η ολοκλήρωση της δημιουργίας ενός δένδρου απόφασης πολλές φορές απαιτεί την αφαίρεση περιττών συγκρίσεων ή και τη διαγραφή ολόκληρων κλαδιών με στόχο την καλύτερη απόδοση της κατηγοριοποίησης. Η φάση του κλαδέματος έχει ως σκοπό τη βέλτιστη απόδοση του δένδρου.

Οι αλγόριθμοι ταξινόμησης που βασίζονται στα δέντρα απόφασης, περιλαμβάνουν δύο διακριτές φάσεις:

1. **Τη φάση οικοδόμησης (building phase):** Σε αυτή την πρώτη φάση, η οποία χρίζει μεγαλύτερης έρευνας και προσπάθειας, το σύνολο των δεδομένων εκπαίδευσης χωρίζεται πολλές φορές, έως ότου όλα τα αντικείμενα σε ένα τμήμα του ανωτέρω συνόλου να ανήκουν στην ίδια κλάση.

2. **Τη φάση κλαδέματος (pruning phase):** Έπειτα, αφού έχει ήδη δημιουργηθεί το δέντρο απόφασης, οι περισσότεροι αλγόριθμοι εκτελούν τη φάση του κλαδέματος, περικόπτοντας κάποιους από τους κόμβους, προκειμένου αφενός να αποτραπούν επικαλύψεις, και αφετέρου το δέντρο να έχει υψηλότερη ακρίβεια ταξινόμησης.

Τα πλεονεκτήματα από τη χρήση δένδρων αποφάσεων κατηγοριοποίησης είναι πολλά και παρατίθενται παρακάτω:

i. Τα δένδρα απόφασης είναι εύκολα στη χρήση και αποτελεσματικά, με κανόνες κατανοητούς και βατούς ως προς την ερμηνεία τους.

ii. Δένδρα απόφασης μπορούν να κατασκευαστούν και για τα δεδομένα με πολλά γνωρίσματα.

iii. Λειτουργούν πάρα πολύ καλά σε μεγάλες βάσεις δεδομένων λόγω του γεγονότος ότι το μέγεθος του δένδρου είναι ανεξάρτητο από το μέγεθος της βάσης.

iv. Η ευρωστία που επιδεικνύουν αναφορικά με το θόρυβο που ενδέχεται να παρουσιαστεί στα δεδομένα που απαρτίζουν το χώρο του προβλήματος.

v. Η ανοχή στην απουσία τιμών (missing values), σε κάποια χαρακτηριστικά του σώματος εκπαίδευσης.

vi. Η χρήση ακόμα και συνεχών (μη διακριτών) χαρακτηριστικών και η προσέγγιση μη διακριτών συναρτήσεων στόχου, μέσω εξειδικευμένων τεχνικών που αναλαμβάνουν τη διακριτοποίηση τους (discretization), τη διαδικασία δηλαδή της μετατροπής συνεχών αριθμητικών χαρακτηριστικών σε κατηγορικά.

vii. Η δυνατότητα μεταφοράς του παραγόμενου μοντέλου από δένδρο απόφασης σε ένα σύνολο κανόνων, προς διευκόλυνση της κατανόησής του.

Δε λείπουν ωστόσο και τα μειονεκτήματα από τη χρήση δένδρων απόφασης μερικά από τα οποία είναι:

i. Δε χειρίζονται εύκολα δεδομένα, τα γνωρίσματα των οποίων αποτελούνται από συνεχείς τιμές.

ii. Υπάρχει η πιθανότητα υπερ-προσπαρμογής ενός δένδρου στα σύνολα δεδομένων εκπαίδευσης.

1.5.1.2 ΝΕΥΡΩΝΙΚΑ ΔΙΚΤΥΑ

Πέρα από τις μεθόδους ταξινόμησης που βασίζονται στα δέντρα και τους κανόνες απόφασης, τα τεχνητά νευρωνικά δίκτυα (artificial neural networks) είναι επίσης μια διαδεδομένη μέθοδος ταξινόμησης (Michie et al, 1995, Kotsiantis, 2007).

Συγκεκριμένα, είναι μια δομή που αποτελείται από ένα δίκτυο νευρώνων (neurons) οι οποίοι συνδέονται μεταξύ τους και αποτελούν τα δομικά στοιχεία του δικτύου. Κάθε τέτοιος κόμβος δέχεται ένα σύνολο αριθμητικών εισόδων από διαφορετικές πηγές (είτε από άλλους νευρώνες, είτε από το περιβάλλον), επιτελεί έναν υπολογισμό με βάση αυτές τις εισόδους και παράγει μία έξοδο. Η εν λόγω έξοδος είτε κατευθύνεται στο περιβάλλον, είτε τροφοδοτείται ως είσοδος σε άλλους νευρώνες του δικτύου. Η πιο διαδεδομένη κατηγορία νευρωνικών δικτύων είναι τα λεγόμενα δίκτυα πρόσθιας τροφοδότησης (feed-forward neural networks), τα οποία επιτρέπουν την κίνηση των δεδομένων μόνο προς μια κατεύθυνση, δηλαδή από μια

είσοδο προς μια έξοδο και έχουμε και τα δίκτυα που σχηματίζουν κυκλικές δομές τα οποία ονομάζονται ανατροφοδοτούμενα νευρωνικά δίκτυα (recurrent neural networks) (Ρίζος, 1996).

Τα νευρωνικά δίκτυα είναι μία προσέγγιση ανάπτυξης και εκτίμησης μαθηματικών δομών. Οι μέθοδοι αυτοί είναι αποτελέσματα ακαδημαϊκών ερευνών με στόχο την μοντελοποίηση συστημάτων μάθησης. Τα νευρωνικά δίκτυα έχουν την ικανότητα να εξάγουν κάποιο συμπέρασμα από πολύπλοκα ή μη ακριβή δεδομένα και μπορούν να χρησιμοποιηθούν για να εξάγουν πρότυπα και να προσδιορίζουν τάσεις οι οποίες είναι πολύπλοκες για να προσδιοριστούν από ανθρώπους ή από άλλες υπολογιστικές τεχνικές. Ένα εκπαιδευμένο νευρωνικό δίκτυο μπορεί να αντιμετωπιστεί ως ένας ειδικός για την κατηγορία της πληροφορίας που του δόθηκε να αναλύσει. Έτσι μπορεί να χρησιμοποιηθεί για να κάνει κάποιες προβλέψεις, όταν προκύψουν κάποιες νέες περιπτώσεις. Τα νευρωνικά δίκτυα χρησιμοποιούν ένα σύνολο από στοιχεία επεξεργασίας (κόμβους) ανάλογους με τους νευρώνες στο ανθρώπινο μυαλό. Τα στοιχεία αυτά διασυνδέονται μεταξύ τους σε ένα δίκτυο το οποίο μπορεί να αναγνωρίζει πρότυπα μέσα σε ένα σύνολο δεδομένων μόλις αυτά παρουσιαστούν μέσα στα δεδομένα, δηλαδή το δίκτυο μπορεί να μαθαίνει από την εμπειρία όπως ακριβώς κάνουν και οι άνθρωποι. Αυτό διακρίνει τα νευρωνικά δίκτυα από τα παραδοσιακά προγράμματα υπολογιστών, τα οποία απλά ακολουθούν οδηγίες σύμφωνα με μία καλά ορισμένη σειρά.

Το κύριο χαρακτηριστικό των νευρωνικών δικτύων είναι η εγγενής ικανότητα μάθησης. Ως μάθηση μπορεί να οριστεί η σταδιακή βελτίωση της ικανότητας του δικτύου να επιλύει κάποιο πρόβλημα όπως για παράδειγμα η σταδιακή προσέγγιση μίας συνάρτησης. Η μάθηση επιτυγχάνεται μέσω της εκπαίδευσης μιας επαναληπτικής διαδικασίας σταδιακής προσαρμογής των παραμέτρων του δικτύου, σε τιμές κατάλληλες ώστε να επιλύεται με επαρκή επιτυχία το προς εξέταση πρόβλημα. Αφού ένα δίκτυο εκπαιδευτεί, οι παράμετροί του συνήθως παγώνουν στις κατάλληλες τιμές και έπειτα είναι σε λειτουργική κατάσταση. Το ζητούμενο είναι το λειτουργικό δίκτυο να χαρακτηρίζεται από μία ικανότητα γενίκευσης. Αυτό σημαίνει ότι πρέπει να δίνει ορθές εξόδους για εισόδους καινοφανείς και διαφορετικές από αυτές με τις οποίες εκπαιδεύτηκε.

Σε ένα νευρωνικό δίκτυο πρόσθιας τροφοδότησης, τα κύρια βήματα για την κατασκευή ενός μοντέλου ταξινόμησης, είναι τα εξής (Aggarwal & Yu, 1999, Βαζιργιάννης & Χαλκίδη, 2003):

- Η αναγνώριση των χαρακτηριστικών εισόδου και εξόδου
- Η κατασκευή ενός δικτύου με την κατάλληλη τοπολογία
- Η επιλογή του σωστού συνόλου εκπαίδευσης το οποίο περιλαμβάνει δεδομένα που είναι ορισμένα ανά ζεύγη
- Η εκπαίδευση του δικτύου στην οποία τα δεδομένα εισέρχονται στο νευρωνικό δίκτυο ένα ένα. Το νευρωνικό δίκτυο μαθαίνει συγκρίνοντας τα αποτελέσματα ταξινόμησης ενός αντικειμένου με την γνωστή πραγματική ταξινόμηση αυτού. Τα λάθη από την αρχική ταξινόμηση του πρώτου αντικειμένου χρησιμοποιούνται για να διορθωθεί το δίκτυο μέσω της τροποποίησης των συναρτήσεων των νευρώνων. Η παραπάνω

διαδικασία είναι επαναληπτική. Η επαναληπτική φύση ωστόσο της διαδικασίας εκπαίδευσης σημαίνει ότι ένα νευρωνικό δίκτυο είναι αρκετά αργό.

- Ο έλεγχος του δικτύου χρησιμοποιώντας ένα σύνολο ελέγχου, το οποίο είναι ανεξάρτητο από το σύνολο εκπαίδευσης.

Οι νευρώνες ενός δικτύου χωρίζονται σε τρείς βασικές κατηγορίες:

1) **Τους νευρώνες εισόδου (input neurons):** οι οποίοι δέχονται τις πληροφορίες που θα υποστούν επεξεργασία

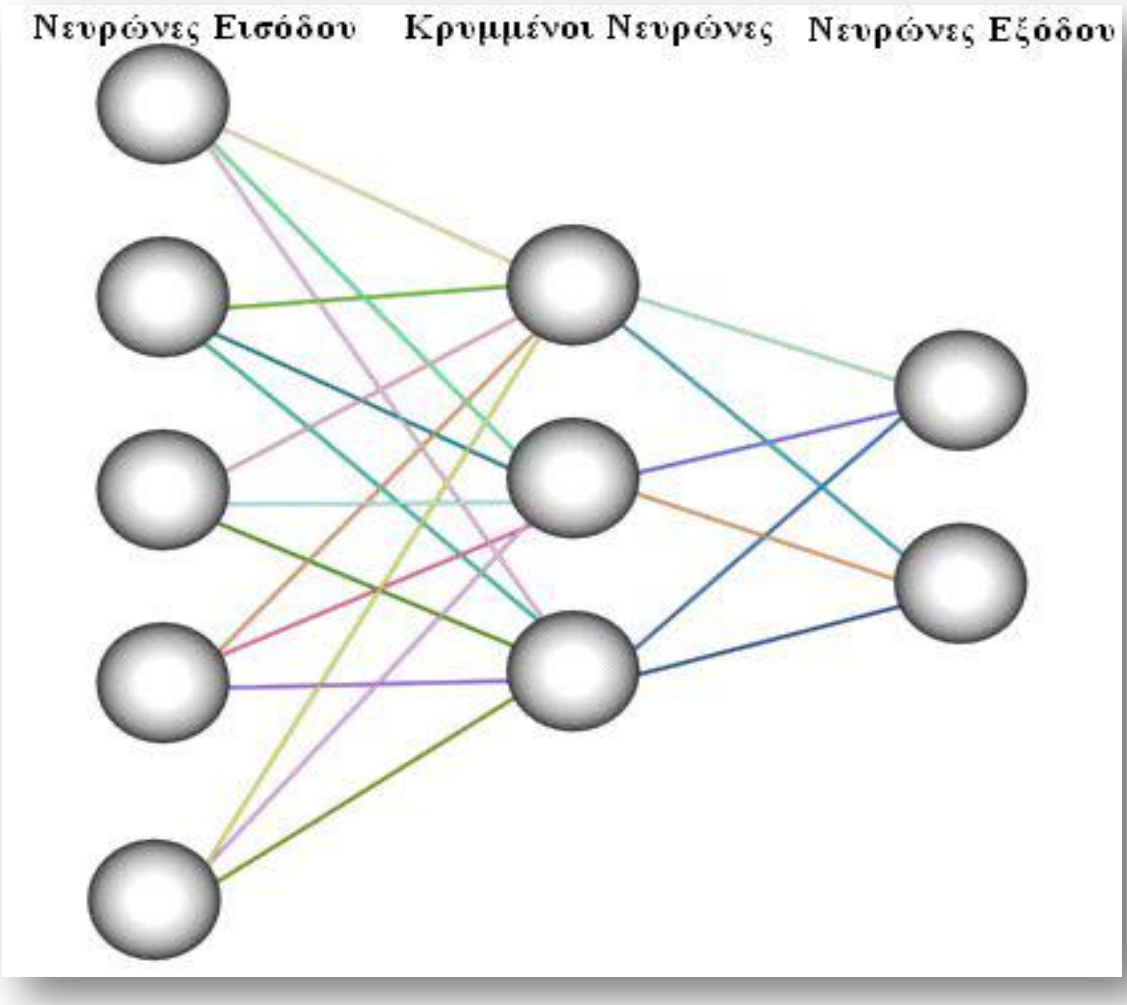
2) **Τους νευρώνες εξόδου (output neurons):** στους οποίους καταλήγουν τα αποτελέσματα της παραπάνω επεξεργασίας

3) **Τους ενδιάμεσους νευρώνες:** οι οποίοι βρίσκονται μεταξύ των νευρώνων εισόδου και εξόδου. Οι τελευταίοι εναλλακτικά ονομάζονται και κρυφοί νευρώνες (hidden neurons).

Ουσιαστικά, οι νευρώνες σε ένα δίκτυο είναι αφενός ένα σύνολο εισερχόμενων τιμών και των αντίστοιχων βαρών τους και αφετέρου μια συνάρτηση που αθροίζει τα παραπάνω βάρη, αντιστοιχώντας τα αποτελέσματα σε ένα νευρώνα εξόδου (Aggarwal & Yu, 1999).

Καταλήγοντας αξίζει να σημειώσουμε ότι εν πολλοίς η εκπαίδευση ενός νευρωνικού δικτύου βασίζεται στον υπολογισμό των τιμών των βαρών που προ αναφέρθηκαν. Ο πιο γνωστός αλγόριθμος, μεταξύ άλλων (Neocleous & Schizas, 2002), στον οποίο βασίζεται ο παραπάνω υπολογισμός, είναι ο αλγόριθμος ανάστροφης μετάδοσης (back propagation algorithm) (Rumelhart et al, 1986). Άλλες προσεγγίσεις που χρησιμοποιούνται για την εκπαίδευση των νευρωνικών δικτύων, με κύριο στόχο την βελτίωση των χρονικών τους επιδόσεων, είναι αυτές των Weigend et al (1990) και Yam & Chow (2001). Επιπλέον, για την εκπαίδευση των νευρωνικών δικτύων μπορούν να χρησιμοποιηθούν τόσο γενετικοί αλγόριθμοι όσο και στατιστικές μέθοδοι.

Στο παρακάτω σχήμα παραθέτουμε μία χαρακτηριστική απεικόνιση ενός νευρωνικού δικτύου, όπου διακρίνονται οι νευρώνες εισόδου, οι κρυμμένοι νευρώνες και οι νευρώνες εξόδου όπως περιγράφτηκαν παραπάνω.



1.5.1.3 ΜΕΘΟΔΟΙ ΠΟΥ ΣΤΗΡΙΖΟΝΤΑΙ ΣΤΟΥΣ ΚΑΝΟΝΕΣ ΑΠΟΦΑΣΗΣ

Στη παρούσα εργασία δε θα αναλύσουμε τους αλγόριθμους που εξάγουν κανόνες απόφασης, παρακάτω όμως θα αναφερθούμε γενικά, παρουσιάζοντας τους πιο βασικούς.

Μια πολύ σημαντική ιδιότητα λοιπόν των δέντρων απόφασης, είναι η ικανότητα μετατροπής τους σε ένα σύνολο κανόνων απόφασης (decision rules) (Quinlan, 1987, 1993). Συγκεκριμένα, δημιουργείται ένας ξεχωριστός κανόνας για κάθε μονοπάτι που ξεκινά από την κορυφή του δέντρου και καταλήγει σε ένα φύλλο που αναπαριστά μια κλάση. Επιπλέον, τα περισσότερα από τα άλλα είδη τυποποίησης των εξαγομένων των αλγορίθμων της εξόρυξης δεδομένων, όπως οι λίστες απόφασης (decision lists), τα προς τα κάτω αναπτυσσόμενα σύνολα κανόνων (ripple down rule sets), τα επαγωγικά λογικά προγράμματα (inductive logic programs) ή τα νευρωνικά δίκτυα (neural networks), μπορούν επίσης να μετατραπούν σε κανόνες. Ειδικά για την μετατροπή των τελευταίων σε κανόνες απόφασης, η διεθνής βιβλιογραφία είναι ιδιαίτερα πλούσια (Towell & Shavlik, 1994, Andrews et al, 1995,

Zhou, 2004). Ωστόσο, αξίζει να σημειωθεί ότι οι κανόνες απόφασης μπορούν επιπλέον να εξαχθούν και απ' ευθείας από το σύνολο εκπαίδευσης μιας βάσης δεδομένων, μέσω μιας σειράς αλγορίθμων ταξινόμησης, οι οποίοι βασίζονται στους κανόνες απόφασης (rule-based methods) (Furnkranz, 1999).

Στόχος των παραπάνω αλγορίθμων είναι η εξαγωγή του μικρότερου δυνατού συνόλου κανόνων απόφασης που είναι συνεπές με τα υπό εκπαίδευση δεδομένα. Οι εξαχθέντες κανόνες απόφασης έχουν την γενική μορφή «If A Then B», με το «If» κομμάτι να αποτελεί ένα συνδυασμό ζευγών από τιμές χαρακτηριστικών, αναπαριστώντας τις επαρκείς συνθήκες για την εφαρμογή-ανάθεση της τιμής της κλάσης που περιγράφεται στο «Then» κομμάτι του κανόνα, στο υπό ταξινόμηση αντικείμενο της βάσης δεδομένων. Ένας αλγόριθμος που βασίζεται στους κανόνες απόφασης, πρέπει να παράγει κανόνες οι οποίοι έχουν υψηλές ικανότητες πρόβλεψης και ταυτόχρονα υψηλή αξιοπιστία. Σημαντικό ρόλο σε αυτό διαδραματίζουν συνήθως μηχανισμοί, που είτε καθιστούν πολύ εξειδικευμένους κανόνες πιο γενικούς, σε μια ξεχωριστή φάση κλαδέματός τους (Furnkranz, 1997), είτε σταματούν την διαδικασία εξειδίκευσης των κανόνων μέσω της χρήσης μέτρων ποιότητας. Αυτά τα μέτρα ποιότητας, χρησιμοποιούνται τόσο στην διαδικασία εξαγωγής των κανόνων όσο και στην διαδικασία ταξινόμησης του εκάστοτε αλγορίθμου. Αφενός, στην διαδικασία εξαγωγής των κανόνων, ένα μέτρο αξιολόγησης της ποιότητάς τους μπορεί να χρησιμοποιηθεί σαν κριτήριο της διαδικασίας εξειδίκευσης ή και γενίκευσης των κανόνων, αφετέρου στην διαδικασία της ταξινόμησης, μια τιμή ενός μέτρου αξιολόγησης ποιότητας μπορεί να αντιστοιχιστεί σε κάθε κανόνα, για την επίλυση συγκρούσεων στην περίπτωση που πολλοί κανόνες ταυτόχρονα ικανοποιούν το προς ταξινόμηση αντικείμενο.

Οι An & Cercone (2000) αναφέρονται αναλυτικά στα σημαντικότερα από τα μέτρα αξιολόγησης της ποιότητας των κανόνων. Αξιόλογες αναφορές στα παραπάνω μέτρα γίνονται επίσης, μεταξύ άλλων, και από τους Lavrac et al (1999), Stefanowski & Vanderpooten (2001), Flach & Lavrac (2003) και Tsumoto (2003). Στην διεθνή βιβλιογραφία υπάρχει ένας πολύ μεγάλος αριθμός αλγορίθμων ταξινόμησης που βασίζονται στους κανόνες απόφασης. Αναλυτική αναφορά σε αυτούς γίνεται στον Furnkranz (1999). Ένας από τους σημαντικότερους αλγορίθμους που βασίζεται στους κανόνες απόφασης είναι ο αλγόριθμος RIPPER (Cohen, 1995), ο οποίος διαμορφώνει κανόνες μέσα από μια συνεχή διαδικασία ανάπτυξης (growing) και κλαδέματος (pruning). Στην διάρκεια της πρώτης φάσης, οι δημιουργηθέντες κανόνες είναι πιο συνεπτυγμένοι, με στόχο την καλύτερη δυνατή προσαρμογή τους στα δεδομένα του συνόλου εκπαίδευσης, ενώ στην δεύτερη φάση συμβαίνει ακριβώς το αντίθετο, με στόχο την καλύτερη απόδοση του αλγορίθμου σε νέα δεδομένα. Άλλοι σημαντικοί αλγόριθμοι είναι αυτοί της οικογένειας AQ (Michalski & Chilausky, 1980), ο αλγόριθμος PART (Frank & Witten, 1998) καθώς και ο CN2 (Clark & Niblett, 1989). Ειδικά ο αλγόριθμος CN2 είναι από τους πιο σημαντικούς αλγόριθμους που βασίζονται σε κανόνες. Βασισμένος στην «If A Then B» μορφή των κανόνων, χρησιμοποιεί μια συνάρτηση για τον τερματισμό της διαδικασίας κατασκευής τους, βάσει μιας εκτίμησης για τον θόρυβο που εμπεριέχεται στα δεδομένα. Το εξαγόμενο αποτέλεσμα του CN2 είναι ένα σύνολο διατεταγμένων «If A Then B» κανόνων, γνωστό και ως λίστα απόφασης (decision list) (Rivest, 1987). Αξίζει ακόμα να αναφέρουμε τον αλγόριθμο CL2, ο οποίος εξάγει κανόνες απόφασης χρησιμοποιώντας διαδικασίες ομαδοποίησης.

1.5.2 ΣΥΣΤΑΔΙΟΠΟΙΗΣΗ

Η συσταδιοποίηση ή αλλιώς ομαδοποίηση (clustering) αφορά τον διαχωρισμό (partition) των αντικειμένων μιας βάσης δεδομένων σε μη συνδεδεμένες μεταξύ τους και ομοιογενείς ομάδες, κατά τέτοιο τρόπο ώστε τα αντικείμενα του συνόλου που ανήκουν σε μια ομάδα, να είναι πιο όμοια μεταξύ τους, παρά με τα αντικείμενα που ανήκουν σε διαφορετικές ομάδες (Jain et al, 1999, Larose, 2004). Ένα ιδιαίτερο χαρακτηριστικό της ομαδοποίησης, σε αντίθεση με την κατηγοριοποίηση, είναι ότι η δομή και το πλήθος των ομάδων είναι καταρχάς άγνωστα και καθορίζονται δε από τον εκάστοτε αλγόριθμο συσταδιοποίησης (Zaiane, 1999). Αυτοί οι αλγόριθμοι βασίζονται στο σύνολο τους στην αρχή της μεγιστοποίησης της ομοιότητας ανάμεσα στα αντικείμενα την ίδιας ομάδας (intra-class similarity) και την ταυτόχρονη αρχή της ελαχιστοποίησης της ομοιότητας μεταξύ των αντικειμένων διαφορετικών ομάδων (inter-class similarity). Αξίζει να σημειωθεί ότι η ερμηνεία των ομάδων που προκύπτουν από την ανωτέρω διαδικασία καθορίζεται από τον εκάστοτε χρήστη (Berry & Linoff, 2004).

Από τον παραπάνω ορισμό προκύπτει άμεσα και η βασική διαφορά μεταξύ κατηγοριοποίησης και συσταδιοποίησης. Στην κατηγοριοποίηση ο αριθμός και η ουσία των συστάδων αποτελεί πληροφορία εκ των προτέρων γνωστή. Εξαιτίας αυτού, στη συσταδιοποίηση εφαρμόζεται πάντα μη εποπτευόμενη μάθηση, εν αντιθέση με την κατηγοριοποίηση όπου λόγω της πρότερης γνώσης των κλάσεων κάνουμε χρήση της εποπτευόμενης μάθησης. Στην συσταδιοποίηση δεν υπάρχουν προκαθορισμένες κατηγορίες ομαδοποίησης αλλά οι εγγραφές συγκεντρώνονται σε ομάδες με βάση το κριτήριο που θέτει ο χρήστης για κάθε συστάδα όπως για παράδειγμα, η ομαδοποίηση πελατών που αγοράζουν παρόμοια αγαθά. Σκοπός είναι η δημιουργία συστάδων με όσο το δυνατόν περισσότερα κοινά χαρακτηριστικά εντός της εκάστοτε ομάδας, ενώ ταυτόχρονα η μία ομάδα από την άλλη θα πρέπει να διαφοροποιείται ικανοποιητικά ώστε να μη συγχέονται. Δηλαδή θα πρέπει να δημιουργηθούν διακριτές ομάδες με βάση ξεκάθαρα χαρακτηριστικά που περιγράφουν την κάθε ομάδα και την κάνουν να ξεχωρίζει από τις υπόλοιπες.

Μερικά βασικά ζητήματα που προκύπτουν στην συσταδιοποίηση είναι τα παρακάτω:

- **Ο χειρισμός των ακραίων σημείων:** Πρόκειται στην ουσία για δεδομένα που στην πράξη δεν ανήκουν σε καμία συστάδα. Μπορούν από μόνα τους να θεωρηθούν ως ξεχωριστές συστάδες κάτι που καθιστά κάθε προσπάθεια συσταδιοποίησης φτωχή.

- **Τα δυναμικά δεδομένα:** Αυτά μπορεί να υπάρχουν σε βάσεις δεδομένων και τα οποία καθιστούν και τις ίδιες τις συστάδες δυναμικά μεταβαλλόμενες στο χρόνο.

- **Το είδος των δεδομένων που χρησιμοποιούνται:** Στη προκειμένη περίπτωση δεν είμαστε ακόμα σε θέση να έχουμε καμιά περαιτέρω πληροφορία όσον αφορά τα γνωρίσματα των ομάδων.

- **Η μοναδικότητα της λύσης του προβλήματος:** Πολλές φορές δε γίνεται να επιτευχθεί σε πολλά από τα προβλήματα συσταδιοποίησης, μιας και το ακριβές πλήθος των ομάδων που απαιτούνται δεν είναι και τόσο εύκολο να προσδιοριστεί.

Σε αυτό το σημείο πρέπει να αναφέρουμε ότι σημαντικό ρόλο παίζει και η απόσταση μεταξύ των συστάδων, η οποία θα πρέπει να ορίζεται καταλλήλως. Μια αντιπροσωπευτική λίστα μεθόδων υπολογισμού αποστάσεων είναι αυτή που περιγράφεται παρακάτω:

1) **Απόσταση απλού συνδέσμου:** Η μικρότερη απόσταση μεταξύ δύο στοιχείων των δύο συστάδων

2) **Απόσταση πλήρους συνδέσμου:** Η μεγαλύτερη απόσταση μεταξύ δύο στοιχείων των δύο συστάδων

3) **Μέση απόσταση:** Η μέση απόσταση μεταξύ των στοιχείων των δύο συστάδων

4) **Απόσταση κέντρων βάρους:** Η απόσταση μεταξύ των κέντρων βάρους των δύο συστάδων.

Η συσταδιοποίηση διακρίνεται σε τρεις βασικές μεθόδους:

1. **Μέθοδοι διαχωρισμού (partitioning methods):** Δημιουργούν ομάδες από ένα δεδομένο αρχικό σύνολο αντικειμένων με κάθε ομάδα να αντιπροσωπεύει ένα cluster και να ικανοποιούνται οι εξής δύο συνθήκες: (α) κάθε cluster περιέχει τουλάχιστον ένα αντικείμενο και (β) κάθε αντικείμενο ανήκει σε ένα μόνο cluster.

2. **Ιεραρχικές μέθοδοι (hierarchical methods):** Διασπούν το αρχικό σύνολο δεδομένων δημιουργώντας μια ιεραρχική δομή από clusters και διακρίνονται σε agglomerative (bottom-up) ή divisive (top-down) ανάλογα με τον τρόπο που γίνεται η διάσπαση.

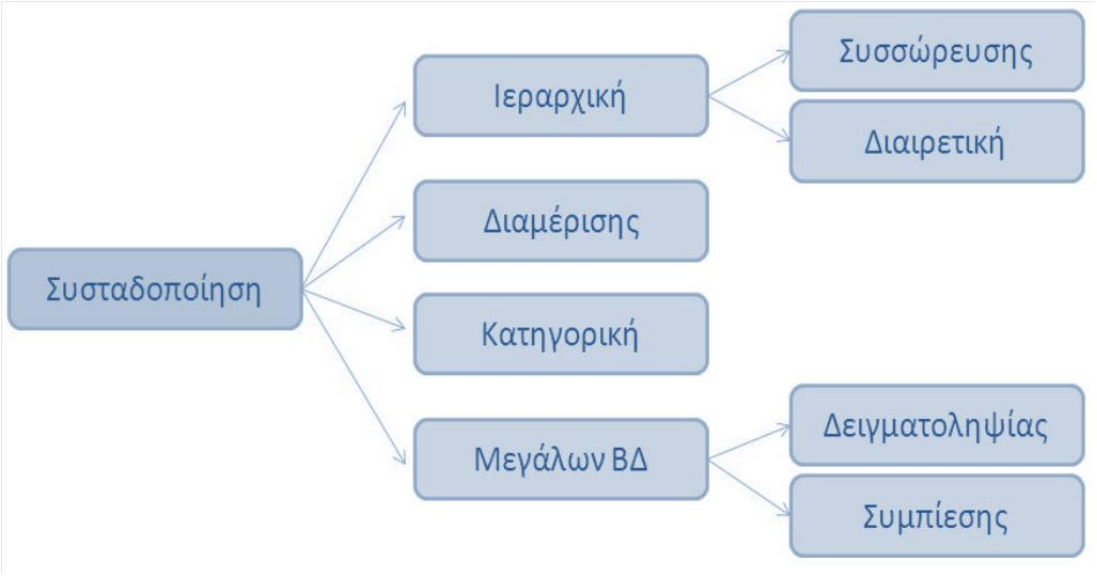
3. **Μέθοδοι βασισμένες σε μοντέλα (model-based methods):** Υποθέτουν ότι καθένα από τα clusters περιγράφεται από ένα μαθηματικό μοντέλο και εντοπίζουν τα αντικείμενα που ανήκουν σε κάθε cluster, ώστε να ικανοποιούν το αντίστοιχο μοντέλο.

Η επιλογή του κριτηρίου για μία σωστή διαδικασία συσταδοποίησης απαιτεί:

- **Επιλογή χαρακτηριστικών γνωρισμάτων:** Ο στόχος είναι να επιλεγούν τα καταλληλότερα γνωρίσματα στα οποία πρόκειται να εφαρμοστεί η συσταδοποίηση ώστε να επιτυγχάνεται η βέλτιστη ομοιογένεια σε κάθε συστάδα. Έτσι η προ-επεξεργασία των δεδομένων πριν την εφαρμογή της διαδικασίας συσταδοποίησης κρίνεται απαραίτητη.

- **Επιλογή αλγορίθμων συσταδοποίησης:** Σε αυτό το στάδιο γίνεται η επιλογή ενός αλγορίθμου που θα οδηγήσει σε ένα καλό σχήμα συσταδοποίησης για ένα σύνολο δεδομένων. Για τη επιλογή του αλγορίθμου χρησιμοποιείται το μέτρο γειτνίασης και το κριτήριο συσταδοποίησης τα οποία ορίζουν απόλυτα τον αλγόριθμο, καθώς επίσης και η δυνατότητά του να καθορίσει ένα σχήμα συσταδοποίησης που να προσαρμόζεται στο συγκεκριμένο σύνολο δεδομένων.

Οι αλγόριθμοι συσταδιοποίησης μπορούν να ταξινομηθούν στις ακόλουθες κατηγορίες:



- **Επικύρωση αποτελεσμάτων:** Σε αυτή τη φάση αξιολογούνται τα αποτελέσματα του αλγορίθμου συσταδοποίησης σύμφωνα με κατάλληλα κριτήρια ορθότητας συσταδοποίησης και τεχνικές. Παράδειγμα ενός τέτοιου κριτηρίου είναι η σύγκριση των αποτελεσμάτων της ανάλυσης με κάποια ήδη γνωστά αποτελέσματα ή η σύγκριση των αποτελεσμάτων δύο διαφορετικών συσταδοποιήσεων. Η ποιότητα της συσταδοποίησης εξαρτάται από την ομοιότητα (δηλαδή μεγάλη ομοιότητα εντός της συστάδας - μικρή ομοιότητα μεταξύ των συστάδων) και την μέθοδο υλοποίησης της συσταδοποίησης.

- **Ερμηνεία των αποτελεσμάτων:** Αποτελεί το τελευταίο στάδιο της διαδικασίας συσταδοποίησης, όπου οι αναλυτές καλούνται να εξάγουν γνώση από τις παραχθείσες συστάδες, συνδυάζοντας κι άλλα στοιχεία, αναλύσεις, με σκοπό το καλύτερο και εγκυρότερο αποτέλεσμα.

Μια μέθοδος συσταδοποίησης είναι καλή αν παράγει συστάδες καλής ποιότητας δηλαδή συστάδες με μεγάλη ομοιότητα εντός της συστάδας.

1.5.3 ΑΝΑΛΥΣΗ ΣΥΣΧΕΤΙΣΗΣ

Η ανάλυση συσχέτισης (association analysis) έχει σαν βασικό της στόχο την ανακάλυψη κρυμμένων συσχετίσεων μεταξύ των χαρακτηριστικών μιας βάσης δεδομένων. Με άλλα λόγια, η παραπάνω ανάλυση ψάχνει να βρει κανόνες για την ποσοτικοποίηση των σχέσεων μεταξύ δύο ή περισσοτέρων χαρακτηριστικών μιας βάσης δεδομένων (Larose, 2004). Οι κανόνες αυτοί ονομάζονται κανόνες συσχέτισης (association rules), και έχουν την μορφή «If A then B» (Agrawal et al, 1996). Οι κανόνες συσχέτισης χαρακτηρίζονται από το κατώφλι στήριξης (support threshold), που αναγνωρίζει τα στοιχεία των βάσεων δεδομένων που εμφανίζονται συχνά σε αυτά, καθώς και το κατώφλι εμπιστοσύνης (confidence threshold), που είναι η υπό συνθήκη πιθανότητα (conditional probability) ένα στοιχείο να εμφανίζεται σε

μια διαδικασία όταν ένα άλλο στοιχείο εμφανίζεται επίσης (Zaiane, 1999). Αξίζει να σημειωθεί ότι η ανάλυση συσχέτισης είναι γνωστή στον επιχειρηματικό κόσμο σαν ανάλυση συνάφειας (affinity analysis) με πολλές εφαρμογές (Berry & Linoff, 2004).

1.5.4 ΠΑΛΙΝΔΡΟΜΗΣΗ

Η παλινδρόμηση (regression) είναι η παλαιότερη και η πλέον γνωστή στατιστική τεχνική που υλοποιείται εντός των πλαισίων της εξόρυξης δεδομένων. Κύριος σκοπός εδώ είναι η πρόβλεψη της τιμής μιας μεταβλητής μελετώντας τις τιμές που είχε στο παρελθόν. Συγκεκριμένα, η παλινδρόμηση χρησιμοποιώντας μια βάση αριθμητικών δεδομένων, αναπτύσσει μια μαθηματική σχέση που ταιριάζει στα δεδομένα αυτά. Στην συνέχεια, η μαθηματική αυτή σχέση χρησιμοποιείται για την πρόβλεψη μελλοντικής συμπεριφοράς, εφαρμόζοντας σε αυτήν νέα αριθμητικά δεδομένα. Ο βασικός περιορισμός της συγκεκριμένης τεχνικής είναι ότι εφαρμόζεται καλά μόνο σε συνεχή ποσοτικά δεδομένα (βάρος, ταχύτητα ή ηλικία). Αντίθετα, η παλινδρόμηση δεν λειτουργεί καλά με κατηγορικά δεδομένα (Fayyad et al, 1996, Draper & Smith, 1997).

1.5.4.1 Η ΕΠΙΔΡΑΣΗ ΤΗΣ ΠΑΛΙΝΔΡΟΜΗΣΗΣ ΚΑΙ ΟΙ ΠΑΡΕΡΜΗΝΕΙΕΣ ΣΤΙΣ ΟΠΟΙΕΣ ΟΔΗΓΕΙ

Η πρώτη προσπάθεια για τη μελέτη της σχέσης μεταξύ δύο μεταβλητών έγινε από τον Sir Francis Galton για την μελέτη της σχέσης του ύψους των παιδιών με τους γονείς τους. Από την μελέτη αυτή προήλθε και ο όρος παλινδρόμηση (regression) που ουσιαστικά αναφέρεται στην παλινδρόμηση προς την κατεύθυνση του μέσου (regression towards the mean). Ο όρος προήλθε από την παρατήρηση του Galton ότι υπάρχει μια τάση όπου ακραίες, ως προς το μέσο τους, παρατηρήσεις της ανεξάρτητης τιμής αντιστοιχούν σε παρατηρήσεις της εξαρτημένης τιμής που δεν είναι το ίδιο ακραίες αλλά είναι πλησιέστερα προς τον μέσο τους. Με απλούστερο τρόπο μπορεί να πει κανείς ότι ακραίες παρατηρήσεις ακολουθούνται από λιγότερο ακραίες παρατηρήσεις δηλαδή αυτές που είναι πλησιέστερα προς το κέντρο. Αυτό κάνει το διάγραμμα σημείων να έχει την μορφή μπάλας του Αμερικάνικου ποδοσφαίρου. Μελετώντας αρχεία για οικογένειες, τα οποία αγόρασε, ο Galton συγκέντρωσε τα ύψη 205 ζευγαριών από γονείς και 928 ενήλικα παιδιά των γονέων αυτών. Δοθέντος ότι το μέσο ύψος των ανδρών είναι, περίπου, 8% μεγαλύτερο από ότι το μέσο ύψος των γυναικών, ο Galton πολλαπλασίασε τα ύψη των γυναικών στο δείγμα του με το συντελεστή 1.08, έτσι ώστε τα ύψη αυτά των γυναικών να γίνουν συγκρίσιμα με τα ύψη των ανδρών του δείγματος. Στη συνέχεια, για το μέσο ύψος κάθε ζευγαριού γονέων, υπολογίστηκε ο μέσος όρος έτσι ώστε να βρεθεί ένα μέσο ύψος γονέων. Τα μέσα ύψη γονέων διαιρέθηκαν στη συνέχεια σε εννέα διαστήματα. Για κάθε κατηγορία μέσου ύψους γονέων υπολογίστηκε το διάμεσο ύψος των παιδιών των γονέων που ανήκαν στην κατηγορία αυτή. Από την μελέτη των δεδομένων ο Galton παρατήρησε ότι, ασυνήθιστα υψηλοί γονείς τείνουν να έχουν παιδιά χαμηλότερα από τους ίδιους ενώ, ασυνήθιστα χαμηλοί γονείς έχουν συνήθως υψηλότερα παιδιά. Το ύψος κάθε ανθρώπου επηρεάζεται από τα γονίδια που κληρονομεί από τους γονείς του. Για διευκόλυνση της παρουσίασης, ας χαρακτηρίσουμε κάποιον ο οποίος κατά τη στιγμή της σύλληψης έχει προβλεπόμενο ύψος ενηλικίωσης με βάση τα γονίδιά του 1.72, ως ένα άτομο γονιδιακού ύψους 1.72. Δεδομένου ότι το ύψος των ανθρώπων επηρεάζεται από την διατροφή, την άσκηση, και άλλους περιβαλλοντικούς παράγοντες, το ύψος που θα έχει κάποιος στην

ενηλικίωσή του δεν θα αντικατοπτρίζει με ένα τέλειο τρόπο την επίδραση των γονιδίων και επομένως δεν θα αποτελεί μία πλήρη επαλήθευση του προβλεφθέντος με βάση τα γονίδιά του ύψος κατά την παιδική ηλικία. Ένα άτομο πραγματικού ύψους 1.75 ίσως είχε ένα γονιδιακά προβλεφθέν ύψος 1.72, με την διαφορά πραγματικού και προβλεφθέντος ύψους οφειλόμενη σε θετική επίδραση περιβαλλοντικών παραγόντων. Αντίθετα, κάποιος με προβλεφθέν γονιδιακό ύψος 1.78 μπορεί να έχει πραγματικό ύψος στην ενηλικίωση 1.75 εξαιτίας αρνητικών επιδράσεων περιβαλλοντικών παραγόντων. Η πρώτη περίπτωση συμβαίνει συχνότερα απ' ότι η δεύτερη γι' αυτό και τα παρατηρούμενα ύψη παιδιών εξαιρετικά υψηλών γονέων αποτελούν, συνήθως, μια υπέρβαση των γονιδιακών υψών των παιδιών αυτών.

Η προηγηθείσα επιχειρηματολογία δεν συνεπάγεται ότι όλοι οι άνθρωποι θα έχουν σε κάποια μελλοντική στιγμή το ίδιο ύψος. Αν συνέβαινε κάτι τέτοιο θα μπορούσε κανείς να αντιστρέψει την επιχειρηματολογία παρατηρώντας ότι πάρα πολύ υψηλοί άνθρωποι έχουν γονείς κάπως χαμηλότερους από αυτούς ενώ πάρα πολύ χαμηλοί άνθρωποι έχουν κάπως υψηλότερους γονείς. Μήπως αυτό συνεπάγεται ότι τα ύψη των ανθρώπων αποκλίνουν; Ούτε το ένα συμβαίνει ούτε το άλλο. Τα ύψη των ανθρώπων ούτε συγκλίνουν ούτε αποκλίνουν. Θα υπάρχουν πάντοτε εξαιρετικά υψηλοί και εξαιρετικά χαμηλοί άνθρωποι. Αυτό που θα πρέπει να αντιληφθούμε είναι ότι τα ύψη των ανθρώπων επηρεάζονται από τυχαίους παράγοντες και ότι, για ανθρώπους που είναι εξαιρετικά υψηλοί οι τυχαίοι παράγοντες επηρέασαν θετικά το ύψος τους και το έκαναν μεγαλύτερο από ότι αναμενόταν με βάση το γονιδίαμα τους. Η παρερμηνεία αυτή είναι μια λανθασμένη συλλογιστική, και οφείλεται στο φαινόμενο της παλινδρόμησης προς την κατεύθυνση του μέσου, (regression towards the mean) είναι δε ακριβώς η παρερμηνεία της προσωρινής φύσης μιας ακραίας παρατήρησης και ο χαρακτηρισμός της ως τάσης. Η κατάσταση που προκύπτει αποδίδεται στην επίδραση της παλινδρόμησης (regression effect).

1.5.4.2 ΠΑΡΑΔΕΙΓΜΑΤΑ ΕΠΙΔΡΑΣΗΣ ΤΗΣ ΠΑΛΙΝΔΡΟΜΗΣΗ ΣΕ ΔΙΑΦΟΡΟΥΣ ΤΟΜΕΙΣ.

1.Τεστ ευφυΐας.

Σε πολλά σχολεία στο εξωτερικό και κυρίως στις Η.Π.Α, διαμορφώνονται προσχολικά προγράμματα για να ενισχύσουν το IQ των παιδιών. Τα παιδιά που συμμετέχουν στο πρόγραμμα κάνουν το IQ τεστ όταν ξεκινούν το πρόγραμμα και το επαναλαμβάνουν όταν ολοκληρώσουν το πρόγραμμα. Και στις δύο περιπτώσεις τα αποτελέσματα είναι γύρω στο 100 με τυπική απόκλιση γύρω στο 15. Τα στοιχεία δείχνουν ότι τέτοια προγράμματα δεν έχουν κάποιο ιδιαίτερο αποτέλεσμα. Μια παρατήρηση όμως που μπορεί να κάνει κάποιος που θα κοιτάξει περισσότερο τα δεδομένα δείχνει κάτι που προκαλεί έκπληξη. Τα παιδιά τα οποία είχαν απόδοση κάτω από το μέσο στο τεστ πριν αρχίσουν το πρόγραμμα πέτυχαν μία μέση βελτίωση περίπου 5 μονάδων στο τεστ που έδωσαν στο τέλος του προγράμματος. Αντιστρόφως όμως, τα παιδιά εκείνα που απέδωσαν πάνω από το μέσο όρο στο αρχικό τεστ έχασαν, κατά μέσο όρο, περίπου 5 μονάδες στο τελικό τεστ. Θα μπορούσε κανείς να οδηγηθεί στο συμπέρασμα ότι το πρόγραμμα αυτό οδηγεί τελικά σε εξισορρόπηση της ευφυΐας των παιδιών; Ή ότι τα ευφυέστερα παιδιά, επειδή παίζουν με παιδιά μικρότερης ευφυΐας, καταλήγουν να οδηγούν τις δύο αυτές κατηγορίες στην ίδια κατάσταση και οι διαφορές να εξαφανίζονται; Φυσικά δεν συμβαίνει τίποτα από αυτά. Και εδώ έχουμε τη χαρακτηριστική περίπτωση του φαινομένου της επίδρασης της παλινδρόμησης (regression effect) σύμφωνα με το οποίο, σε όλες τις περιπτώσεις εξετάσεων, το γκρουπ με τη χαμηλότερη απόδοση σε μια

πρώτη εξέταση, κατά μέσο όρο, θα αποδώσει καλύτερα σε μια δεύτερη εξέταση και το γκρουπ με την υψηλότερη απόδοση, κατά μέσο όρο, θα αποδώσει χαμηλότερα σε μια δεύτερη εξέταση. Η εσφαλμένη αντίληψη ότι η επίδραση της παλινδρόμησης οφείλεται σε κάτι σημαντικό και όχι απλώς στην διάχυση (spread) των παρατηρήσεων γύρω από την γραμμή είναι αυτό που ονομάζεται παρερμηνεία της παλινδρόμησης (regression fallacy).

Μια άλλη διάσταση της παλινδρόμησης προς την κατεύθυνση του μέσου στα τεστ ευφυΐας (IQ tests) είναι η εξής: Σύμφωνα με μία μελέτη που έγινε στην Αμερική παιδιά ηλικία τεσσάρων ετών με IQ 120 συνήθως, όταν ενηλικιωθούν, επιτυγχάνουν σκορ στο IQ τεστ περίπου 110. Παρομοίως, παιδιά τεσσάρων ετών με IQ σκορ 70 έχουν ένα μέσο σκορ στο IQ τεστ όταν ενηλικιωθούν 85. Αυτό δεν συνεπάγεται ότι θα υπάρχουν λιγότεροι ενήλικες απ' ότι παιδιά με πολύ υψηλά ή πολύ χαμηλά αποτελέσματα στο IQ τεστ. Παρότι όσοι άνθρωποι ξεκινούν στην παιδική ηλικία με υψηλό ή χαμηλό IQ σκορ, συνήθως, θα παλινδρομήσουν προς την κατεύθυνση του μέσου, οι θέσεις τους θα παρθούν (θα αντικατασταθούν) από άλλους οι οποίοι στην παιδική τους ηλικία θα έχουν IQ σκορ πλησιέστερα προς τον μέσο.

2.Εκπαίδευση.

Ένα παράδειγμα λανθασμένης ερμηνείας φαινομένων που οφείλονται στην παλινδρόμηση προς την κατεύθυνση του μέσου εμφανίζεται στην αξιολόγηση των φοιτητών. Έχει παρατηρηθεί ότι οι φοιτητές εκείνοι οι οποίοι έχουν τους υψηλότερους βαθμούς στις εξετάσεις προόδου συνήθως, δεν αποδίδουν εξίσου καλά στην τελική εξέταση ενώ, εκείνοι οι οποίοι έχουν χαμηλή βαθμολογία στην εξέταση προόδου, πολλές φορές βελτιώνουν την απόδοσή τους στην τελική εξέταση. Θα μπορούσε αυτό να εκληφθεί ως ένδειξη ότι η απόδοση των φοιτητών συγκλίνει προς μια ανησυχητική μετριότητα με τους ασθενείς φοιτητές να βελτιώνονται και τους καλούς φοιτητές να χειροτερεύουν ή αντιστρέφοντας το προηγούμενο επιχείρημα, το γεγονός ότι αυτοί που πέτυχαν την υψηλότερη βαθμολογία στην τελική εξέταση δεν απέδωσαν εξίσου καλά στην εξέταση προόδου σημαίνει ότι η απόδοση αποκλίνει από τον μέσο; Και στις δύο περιπτώσεις η απάντηση είναι αρνητική. Η εξαιρετικά υψηλή απόδοση σε οποιαδήποτε εξέταση εμπεριέχει και έναν παράγοντα καλής τύχης ενώ η χαμηλή απόδοση έναν παράγοντα ατυχίας. Οι φοιτητές εκείνοι που πέτυχαν την υψηλότερη βαθμολογία σε οποιαδήποτε εξέταση είναι, κυρίως, φοιτητές πάνω από το μέσο όρο που πέτυχαν εξαιρετικά υψηλή βαθμολογία γιατί τα θέματα των εξετάσεων ήταν θέματα που, εξαιτίας της καλής προετοιμασίας τους, είχαν την ευχέρεια να απαντήσουν. Είναι περισσότερο πιθανό ότι οι φοιτητές αυτοί είναι καλοί φοιτητές που απέδωσαν εξαιρετικά καλά από το ενδεχόμενο να ήταν εξαιρετικά καλοί φοιτητές που είχαν μια άσχημη μέρα. Όσοι επιτυγχάνουν τις υψηλότερες βαθμολογίες σε μια οποιαδήποτε εξέταση είναι πολύ πιθανό ότι δεν απέδωσαν εξίσου καλά στην προηγούμενη εξέταση και δεν θα αποδώσουν το ίδιο καλά στην επόμενη εξέταση. Η παλινδρόμηση προς την κατεύθυνση του μέσου μπορεί να θεωρηθεί κι ως μια περίπτωση κακής χρήσης διαθέσιμων δεδομένων. Αν για την αξιολόγηση φοιτητών σε μια εξέταση επιλέξουμε με τυχαίο τρόπο φοιτητές, η μέση βαθμολογία τους θα αποτελεί μια αμερόληπτη εκτίμηση του μέσου του πληθυσμού. Εάν όμως, μετά την εξέταση, ξεχωρίσουμε τους φοιτητές εκείνους που απέδωσαν εξαιρετικά καλά, αυτοί βέβαια δεν αποτελούν ένα τυχαίο δείγμα, αφού έχουν επιλεγεί ακριβώς επειδή είχαν τις υψηλότερες βαθμολογίες. Σε οποιοδήποτε δείγμα οι υψηλότερες τιμές αποτελούν μια υπερεκτίμηση (overestimate) του μέσου του πληθυσμού. Για να έχουμε αμερόληπτες εκτιμήσεις θα πρέπει να έχουμε ένα τυχαίο δείγμα που δεν στηρίζεται στα αποτελέσματα αυτά καθαυτά.

3.Στρατιωτικό.

Ένας εκπαιδευτής πιλότων παρατήρησε ότι πολύ καλές προσγειώσεις συνήθως, ακολουθούνται από προσγειώσεις που δεν είναι εξίσου καλές, ενώ μέτριες προσγειώσεις ακολουθούνται, συνήθως από καλύτερες. Υποπίπτοντας στην λανθασμένη προσέγγιση που οφείλεται στην παρερμηνεία της παλινδρόμησης στην κατεύθυνση του μέσου ο εκπαιδευτής ισχυρίστηκε ότι η ακολουθία αυτή συμβαίνει γιατί συνήθιζε να επαινεί τις καλές προσγειώσεις και να κριτικάρει έντονα τις μέτριες. Για το λόγο αυτό έβγαλε το συμπέρασμα, σε αντίθεση από την κοινά αποδεκτή άποψη με βάση την έρευνα για την μαθησιακή διδασκαλία, ότι ο έπαινος έχει αρνητικά αποτελέσματα στην προσπάθεια ενώ η έντονη κριτική έχει θετικά αποτελέσματα.

4.Οικονομικό:

Ένα χαρακτηριστικό παράδειγμα του προβλήματος στον τομέα των οικονομικών δίνεται στο βιβλίο με τον προκλητικό τίτλο Ο Θρίαμβος της Μετριότητας στις Επιχειρήσεις (The Triumph of Mediocrity in Business). Ο συγγραφέας ανακάλυψε ότι επιχειρήσεις με εξαιρετικά υψηλά κέρδη σε κάθε δεδομένη χρονιά έχουν χαμηλότερα κέρδη την επόμενη χρονιά ενώ επιχειρήσεις με πολύ χαμηλά κέρδη, εν γένει επιτυγχάνουν καλύτερα αποτελέσματα το επόμενο έτος. Με αυτές τις ενδείξεις κατέληξε στο συμπέρασμα ότι οι ισχυρές επιχειρήσεις γίνονται ασθενέστερες ενώ οι ασθενείς γίνονται ισχυρότερες με αποτέλεσμα σύντομα να γίνουν όλες οι επιχειρήσεις μεσαίου μεγέθους. Η τελείως λανθασμένη προσέγγιση του συγγραφέα είναι προφανής. Ο διάσημος στατιστικός Harold Hotelling εξήγησε το λάθος αυτό ως εξής: «Οι αποδόσεις των επιχειρήσεων με ακραίες αποδόσεις τείνουν, συχνά, προς την κατεύθυνση του κέντρου ενώ εκείνες με μεσαίες αποδόσεις σε ένα σύνολο τείνουν προς τα άκρα. Μερικές, βελτιώνουν την απόδοσή τους ενώ άλλες χειροτερεύουν. Ο μέσος των κερδών του αρχικού συνόλου των επιχειρήσεων που βρισκόταν στο κέντρο είναι ενδεχόμενο, επομένως, να επιδείξει κάποια μικρή μεταβολή δοθέντος ότι, θετικές και αρνητικές αποκλίσεις ακυρώνονται στην διαδικασία υπολογισμού του μέσου, ενώ για ένα σύνολο με ακραίες αποδόσεις η μόνη δυνατή κίνηση είναι προς την κατεύθυνση του κέντρου».

ΚΕΦΑΛΑΙΟ 2
ΕΦΑΡΜΟΓΕΣ ΤΗΣ ΕΞΟΡΥΞΗΣ ΓΝΩΣΗΣ ΚΑΙ ΔΕΔΟΜΕΝΩΝ

2.1 ΠΟΥ ΕΦΑΡΜΟΖΕΤΑΙ Η ΕΞΟΡΥΞΗ ΓΝΩΣΗΣ

Τα παραδείγματα εξόρυξης γνώσης από δεδομένα ποικίλουν ανάλογα με τον τομέα στον οποίο εφαρμόζονται. Στη σημερινή εποχή όπου τα δεδομένα υπάρχουν σχεδόν παντού και τις περισσότερες φορές βρίσκονται σε ηλεκτρονική μορφή, η σωστή ανάλυση τους οδηγεί πάντα στην ανάδειξη και οργάνωση της πληροφορίας, η γνώση της οποίας είναι ο σημαντικότερος παράγοντας για την εύρεση μιας στρατηγικής και την ορθολογική λήψη αποφάσεων.

Ο χρηματοοικονομικός τομέας, ο τομέας των τηλεπικοινωνιών, της υγείας και της εκπαίδευσης, ο δημόσιος τομέας καθώς επίσης και αυτός της βιομηχανίας και της έρευνας, αποτελούν ίσως το μεγαλύτερο δείγμα εφαρμογών των τεχνολογιών εξόρυξης γνώσης από δεδομένα.

Στον χρηματοοικονομικό τομέα η παρουσία αρκετών τραπεζικών και ασφαλιστικών ιδρυμάτων σε συνδυασμό με το υπάρχον οικονομικό κλίμα έχει αυξήσει κατά πολύ τον ανταγωνισμό μεταξύ των επιχειρήσεων. Πολλά από τα επιχειρηματικά ζητήματα του κλάδου, όπως η προσέλκυση νέων πελατών που θα αποφέρουν κέρδος, η προώθηση και πώληση επιπρόσθετων προϊόντων ή παροχή υπηρεσιών, η διατήρηση πελατών, ο προσδιορισμός οικονομικού δόλου, η ανάλυση του πιστωτικού κινδύνου ενδεχόμενων πελατών, μπορούν να αντιμετωπιστούν με τα κατάλληλα εργαλεία της εξόρυξης γνώσης.

Στον κλάδο των τηλεπικοινωνιών το διαρκώς μεταβαλλόμενο επιχειρηματικό περιβάλλον με τον αστείρευτο ανταγωνισμό, αναγκάζουν τις εταιρείες να προβούν στην αναζήτηση νέων τρόπων ενίσχυσης της θέσης τους έναντι άλλων. Έχοντας γίνει πλήρως αντιληπτό από τη πλευρά τους το συγκριτικό πλεονέκτημα που τους προσφέρει η εξόρυξη γνώσης από δεδομένα, εφαρμόζουν μεταξύ άλλων τεχνικές έγκαιρης πρόβλεψης διακοπής υπηρεσιών από πελάτες, κατηγοριοποίηση των απαιτήσεων τους, ομαδοποίηση των συνηθειών τους και όλα αυτά με τελικό σκοπό την συγκράτηση των πελατών που ήδη έχουν αλλά και την προσέλκυση νέων.

Ο τομές του λιανικού εμπορίου είναι ένας άλλος κλάδος ιδιαίτερα ανταγωνιστικός, όπου οι εφαρμογές εξόρυξης γνώσης βρίσκουν μεγάλη ανταπόκριση. Οι συνεχείς αλλαγές των καταναλωτικών προτιμήσεων και οι τεράστιοι όγκοι δεδομένων πωλήσεων, κρύβουν πολύτιμα στοιχεία εκ των οποίων ελάχιστα μπορούν να αξιοποιηθούν από τα συμβατικά συστήματα ανάλυσης πληροφορίας. Αντιθέτως οι εφαρμογές εξόρυξης γνώσης δίνουν μια νέα διάσταση στην παλαιότερη και βασικότερη επιχειρηματική διαδικασία που είχε σαν αρχή: «αναλύοντας ότι έγινε στο παρελθόν και κατανοώντας τα αποτελέσματα μπορούμε να γίνουμε αποτελεσματικότεροι στο μέλλον.» Επιπλέον, οι εφαρμογές εξόρυξης γνώσης κάνουν εφικτή μια προσωποποιημένη σχέση με κάθε ένα πελάτη χωριστά, κάτι που εξασφαλίζει την διαχρονική σχέση και την μεγιστοποίηση του κέρδους ανά πελάτη.

Οι επαγγελματίες στο χώρο της υγείας, πάντα αντιμετωπίζουν την ανάγκη να συλλέγουν, να αποθηκεύουν και να αναλύουν μεγάλες ποσότητες δεδομένων που μπορεί να περιλαμβάνουν καρτέλες ασθενών, δοκιμές νέων φαρμάκων, εξάρσεις ασθενειών και πολλά άλλα.

Από όλα τα παραπάνω, συμπεραίνουμε ότι η εξόρυξη γνώσης είναι ένα απαραίτητο εργαλείο σε πολλούς τομείς της σύγχρονης κοινωνίας. Η ραγδαία αύξηση του όγκου δεδομένων έχει καταστήσει σαφές ότι παλιές παραδοσιακές τεχνικές και μέθοδοι, δε μπορούν πλέον να βοηθήσουν στην ανάλυση και οργάνωση της πληροφορίας. Πολύ περισσότερο, δε μπορούν να φέρουν στην επιφάνεια γνώση που τα δεδομένα περιέχουν καλά κρυμμένη και η οποία απαιτεί εφαρμογή ειδικών για να αποκαλυφθεί.

2.2 ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΟΙΚΟΝΟΜΙΑ

Ένας από τους τομείς που εφαρμόζεται κατά κόρον η εξόρυξη δεδομένων είναι αυτός της οικονομίας. Τα οικονομικά δεδομένα συλλέγονται κυρίως από τράπεζες, σουπερμάρκετ και από άλλους οικονομικούς οργανισμούς. Τα δεδομένα αυτά συνήθως είναι αξιόπιστα, ολοκληρωμένα, έχουν υψηλή ποιότητα και απαιτούν συστηματική μέθοδο για την ανάλυση τους. Η συνεισφορά της εξόρυξης δεδομένων στην επιστήμη της οικονομίας συναντάται στην συλλογή, κατανόηση και βελτίωση των δεδομένων, στην δημιουργία και εκτίμηση ενός μοντέλου και στην ανάπτυξη αυτού. Η σωστή ανάλυση των οικονομικών δεδομένων διευκολύνει στο να παρθούν καλύτερες αποφάσεις ενεργώντας σύμφωνα με την ανάλυση της αγοράς. Τα εργαλεία και οι τεχνικές της εξόρυξης δεδομένων βοηθούν στο να αναλύσουμε τα οικονομικά δεδομένα και είναι τέτοια η συμβολή τους έτσι ώστε για παράδειγμα, τα οικονομικά ινστιτούτα να αναγνωρίζουν τις απάτες από παραποιημένα δεδομένα από τις διάφορες βάσεις δεδομένων και από το ιστορικό συναλλαγών που έγιναν από τους πελάτες. Οι τεχνικές οπτικοποίησης βοηθούν στην παρουσίαση δεδομένων με διαφορετικές μορφές, όπως γράφοι που βασίζονται σε συγκεκριμένα γνωρίσματα. Παραδείγματος χάρη προβάλλοντας τα δεδομένα από διάφορες οπτικές γωνίες, μία τράπεζα δύναται να διακρίνει τους πελάτες που έχουν επιχειρήσει παράνομες πράξεις και μετά μια λεπτομερής έρευνα αυτών των ύποπτων περιπτώσεων βοηθάει στην εξιχνίαση των απατών και των εγκλημάτων.

2.3 ΕΦΑΡΜΟΓΕΣ ΕΞΟΡΥΞΗΣ ΓΝΩΣΗΣ ΣΕ ΕΠΙΣΤΗΜΟΝΙΚΑ ΠΕΔΙΑ

Στο σημείο αυτό θα αναφέρουμε συνοπτικά ορισμένες εφαρμογές των μεθόδων της εξόρυξης γνώσης οι οποίες έχουν συλλεχθεί από διάφορα επιστημονικά πεδία.

- **Μάρκετινγκ μέσω mail (Direct mail marketing):** Το Body Shop International δοκιμάζει τεχνικές εξόρυξης γνώσης ώστε να καταφέρει να αυξήσει την αποτελεσματικότητα των παραγγελιών του μέσω email. Τα διοικητικά στελέχη ενδιαφέρονται στο να μειώσουν το κόστος αποστολής διαφημιστικών καταλόγων, εστιάζοντας μόνο σε πελάτες οι οποίοι θεωρούνται ως κερδοφόροι.

- **Κατηγοριοποίηση μέσω μάνατζμεντ και έλεγχος για κάτι νέο (Category management and inventory control):** Η εταιρεία Rubbermaid χρησιμοποιεί μεθόδους εξόρυξης γνώσης για να κατηγοριοποιεί το στυλ της στρατηγικής που χρησιμοποιεί ανάλογα με την αγορά στην οποία εστιάζεται, στο να αγοράζει και να πουλά τα οικονομικά αγαθά που παράγει. Ανάλυση καλαθιού της νοικοκυράς (Market basket analysis): Τα στελέχη της J.Crew Group συνδυάζουν click system analysis μέσα από το επίσημο site τους σε συνδυασμό με την μέθοδο point-of-sale (POS) στις

λιανικές τους πωλήσεις. Με αυτό τον τρόπο θέλουν να δούνε τι ρούχα, τι υποδήματα και άλλα αξεσουάρ αγοράζονται μαζί. Τα δεδομένα μετά θα αναλυθούν και έπειτα θα σταλθούν κατάλογοι με οικονομικά αγαθά και προσφορές των σε on-line αγοραστές.

- **Κατανόηση του προφίλ του κάθε πελάτη (Customer relationship management):** Σύμφωνα με διάφορες εταιρείες λιανικής, άλλες από τον τραπεζικό τομέα είναι δυνατό χρησιμοποιώντας τεχνικές εξόρυξης γνώσης, όπως κανόνες ταξινόμησης ή και ομαδοποίησης να κατατάσσουμε τους πελάτες ανάλογα με προσωπικά τους στοιχεία και την προηγούμενη τους συμπεριφορά προς την εταιρεία που μας ενδιαφέρει σε καλούς, μέτριους και κακούς. Έτσι δίνεται η δυνατότητα να επιλέγουν το αγοραστικό κοινό με το οποίο συναλλάσσονται και στο οποίο εστιάζουν την πολιτική της εταιρείας.

- **Αστρονομία (Astronomy):** Έχει κατασκευαστεί ένα σύστημα με την ονομασία SKICAT από το JPL/Caltech και χρησιμοποιείται από τους αστρονόμους στο να αναγνωρίζουν αυτόματα τους διάφορους γαλαξίες και αστεροειδείς σε μία μεγάλη κλίμακα η οποία περιέχει διάφορα αστρονομικά μεγέθη.

- **Βιολογία (Biology):** Διάφορα συστήματα έχουν κατασκευαστεί στο να εξάγουν κανόνες που αφορούν την δομή των οργανισμών, την ανάλυση του DNA, καθώς και την δυνατότητα για εύρεση φαρμάκων για την καταπολέμηση ασθενειών.

- **Παγκόσμιο μοντέλο κλιματολογικών συνθηκών (Global climate modeling):** Διάφορα συστήματα έχουν υιοθετηθεί τα οποία επιτρέπουν την ανάλυση κλιματολογικών συνθηκών δίνοντας έτσι την δυνατότητα προβλέψεων κλιματολογικών φαινομένων όπως οι κυκλώνες, οι καταιγίδες, οι καύσωνες και άλλα πολλά.

- **Η μέθοδος εξόρυξης από δεδομένα σε οικονομικές εφαρμογές (Data mining for financial applications):** Πολλές φορές τεχνικές της μεθόδου εξόρυξης από δεδομένα όπως τα νευρωνικά δίκτυα και τα δένδρα αποφάσεων μπορούν να χρησιμοποιηθούν από οικονομικούς αναλυτές για την λήψη στρατηγικών αποφάσεων στο ανάλογο οικονομικό πεδίο που ενδιαφέρει κάθε φορά. Φυσικά χρειάζεται και το ανάλογο υπόβαθρο από ιστορικά δεδομένα ώστε να είναι δυνατή η ανάλυση που θα γίνει.

2.4 ΕΞΌΡΥΞΗ ΓΝΩΣΗΣ ΚΑΙ ΠΙΣΤΩΤΙΚΕΣ ΚΑΡΤΕΣ

Οι πιστωτικές κάρτες χρησιμοποιούνται στην αγορά, ως ένα συμπληρωματικό μέσο, για την ολοκλήρωση ενός μεγάλου μέρους των συναλλαγών των καταναλωτών. Αλλιώς μπορούν να χαρακτηριστούν και ως πλαστικό χρήμα. Αν και στις ηλεκτρονικές δραστηριότητες μπορεί να εμφανιστούν ποικίλα προβλήματα, οι πιστωτικές κάρτες παρέχουν ασφάλεια και επιπλέον δεν απαιτούν τη μεταφορά μετρητών. Η χρήση πιστωτικής κάρτας χρησιμεύει σε αγορές αγαθών, των οποίων η εξόφληση μπορεί να πραγματοποιηθεί με δόσεις ενώ ταυτόχρονα μπορούν να καλύψουν και ανάγκες ανάληψης μετρητών. Αντιλαμβανόμαστε λοιπόν ότι οι τράπεζες, για να τις χορηγήσουν, εξετάζουν κάποια κριτήρια αξιολόγησης. Πιο συγκεκριμένα λαμβάνουν υπόψη κυρίως το ατομικό εισόδημα, όπως αυτό εμφανίζεται στο εκκαθαριστικό σημείωμα της εφορίας και ταυτόχρονα αν υπάρχουν ή μη δυσμενή στοιχεία στο σύστημα γνωστό και ως «Τειρεσίας». Μεταξύ των άλλων εξετάζουν το επάγγελμα των αιτούντων, αν είναι μόνιμοι

κάτοικοι της Ελλάδας, καθώς και αν η μορφή απασχόλησης τους είναι ολική ή μερική. Σημαντικό ρόλο κατέχει και το χρονικό διάστημα που εργάζονται, δηλαδή αν οι αιτούντες εργάζονται μόνο λίγους μήνες ή αρκετά έτη. Ο υποψήφιος κάτοχος πιστωτικής κάρτας είναι αναγκαίο να συμπληρώσει την ειδική αίτηση για την χορήγηση της, η οποία αποτελεί και την σύμβαση του με την τράπεζα, στην περίπτωση που εγκριθεί το αίτημα του.

Αξιοσημείωτο είναι ότι η τράπεζα ανάλογα με τα κριτήρια της, μπορεί να χορηγήσει απλή ή χρυσή κάρτα. Καταλαβαίνουμε πως η χρυσή κάρτα απευθύνεται σε άτομα με υψηλά εισοδήματα και υψηλή πιστοληπτική ικανότητα. Παράλληλα είναι φυσιολογικό στις περιπτώσεις καλών πελατών, δηλαδή σε πελάτες που δεν παρουσιάζουν αρνητικά στοιχεία στο σύστημα «Τειρεσίας», με ικανοποιητικό δηλωθέν εισόδημα, και που εξοφλούν έγκαιρα τις δόσεις τους, οι τράπεζες να αναπροσαρμόζουν άμεσα την αύξηση των πιστωτικών ορίων των καρτών. Πάντως είναι απαραίτητο να αναφέρουμε, σύμφωνα με έρευνα, ότι ένας από τους σημαντικότερους λόγους που δεν έχει αναπτυχθεί ιδιαίτερα στην χώρα μας η χρήση των πιστωτικών καρτών οφείλεται στο τραπεζικό σύστημα, που δεν έχει τελειοποιήσει τους μηχανισμούς εξακρίβωσης της πιστοληπτικής ικανότητας των δανειζόμενων. Αυτό το κενό μπορεί να καλύψει η εξόρυξη γνώσης.

2.4.1 ΔΙΑΔΙΚΑΣΙΑ ΕΚΔΟΣΗΣ ΠΙΣΤΩΤΙΚΗΣ ΚΑΡΤΑΣ

Ο υποψήφιος κάτοχος πιστωτικής κάρτας ζητάει από την εκδότρια τράπεζα, που έχει λογαριασμό την έκδοση μιας κάρτας. Αυτό συμβαίνει διότι η κάρτα που εκδίδεται από την ελληνική τράπεζα πρέπει να είναι συνδεδεμένη με κάποιον από τους παγκόσμιους οργανισμούς πιστωτικών καρτών, για να δίνεται ταυτόχρονα η δυνατότητα στον κάτοχο της να μπορεί να την χρησιμοποιεί παντού σε παγκόσμια κλίμακα Η κάρτα εκδίδεται στο όνομα του πελάτη και οι συναλλαγές χρεώνονται σε έναν ανοιχτό λογαριασμό, στον οποίο έχει καθοριστεί το πιστωτικό όριο. Αντιλαμβανόμαστε ότι το πιστωτικό όριο εξαρτάται από την οικονομική επιφάνεια του πελάτη, καθώς και από την πολιτική της τράπεζας που εκδίδει την πιστωτική κάρτα. Ο τρόπος που θα την χρησιμοποιεί ο κάτοχος της μπορεί να την μετατρέψει σε χρήσιμο εργαλείο. Άρα ο χρήστης της δεν πρέπει να υπερβαίνει το πιστωτικό όριο που του δίνει η τράπεζα και συγχρόνως θα πρέπει να εξοφλεί ολόκληρο το οφειλόμενο ποσό έτσι, ώστε να μην χρεώνεται με υπέρμετρους τόκους καθυστέρησης. Υπάρχουν πολλά εμπορικά καταστήματα που προσφέρουν πολλές διευκολύνσεις στους κατόχους των πιστωτικών καρτών, όπως τη δυνατότητα εξόφλησης σε αρκετές άτοκες δόσεις. Επομένως όποιος προτιμά αυτόν τον τρόπο για να κάνει τις αγορές του εξασφαλίζει μια άτοκη πίστωση. Εκτός των άλλων είναι απαραίτητο να γνωρίζουμε ότι οι πιστωτικές κάρτες έχουν όλες τις συναλλακτικές δυνατότητες, ενώ οι αναλήψεις ορίζονται έως κάποιο συγκεκριμένο όριο πάντα μετά από συμφωνία με την τράπεζα. Συνήθως υπάρχει και άτοκη περίοδος χάριτος για αναλήψεις, μέχρι 60 ημέρες. Βεβαίως ορισμένες τράπεζες δίνουν τη δυνατότητα στον πελάτη τους για ανάληψη μετρητών προκαταβολικά, σε περιπτώσεις εκτάκτου ανάγκης, αλλά και την έκδοση κάρτας και σε μέλη της οικογένειας εντός του εγκεκριμένου ποσού. Επίσης διαφοροποιούνται από τα προσωπικά καταναλωτικά δάνεια ως προς τα επιτόκια δανεισμού, ενώ συνοδεύονται με διάφορες παροχές όπως ειδικά προγράμματα εκπτώσεων, ταξιδιωτική ασφάλιση και δώρα. Στην περίπτωση που ο χρήστης της πιστωτικής κάρτας δεν μπορεί να εργαστεί λόγω ατυχήματος είτε ασθένειας τότε διευκολύνεται στην αποπληρωμή της ελάχιστης μηνιαίας καταβολής του. Ο λογαριασμός του κατόχου της πιστώνεται με το 10% της συνολικής οφειλής κάθε μήνα ενώ παράλληλα η πίστωση της ελάχιστης μηνιαίας

καταβολής συνεχίζεται για το χρονικό διάστημα που αυτός δεν μπορεί να εργαστεί, με μέγιστο χρονικό όριο τους 10 μήνες.

2.4.2 ΕΠΙΧΕΙΡΗΜΑΤΙΚΕΣ ΠΙΣΤΩΤΙΚΕΣ ΚΑΡΤΕΣ

Ιδιαίτερο ενδιαφέρον έχουν οι επιχειρηματικές πιστωτικές κάρτες ή διαφορετικά co-branding credit cards. Οι co-branded πιστωτικές κάρτες εκδίδονται από μεγάλα εμπορικά καταστήματα, ασφαλιστικές εταιρίες ή επιχειρήσεις σε συνεργασία με τράπεζες και απευθύνονται σε εξειδικευμένες ομάδες καταναλωτών. Σίγουρα και τα τρίτα μέρη, δηλαδή η επιχείρηση, η τράπεζα, ο καταναλωτής έχουν την δυνατότητα να κερδίσουν. Και αυτό γιατί οι καταναλωτές μπορούν να λαμβάνουν περισσότερες προσθετές παροχές, όπως εκπτώσεις, με το ίδιο σχεδόν ή και μικρότερο επιτόκιο από αυτό των κυρίως πιστωτικών καρτών, οι επιχειρήσεις μεγαλώνουν τον κύκλο της πελατείας τους και κατά συνέπεια αυξάνουν τα έσοδα τους, εξυπηρετούν και προσέχουν καλύτερα τους πελάτες τους ενώ οι τράπεζες γιατί αυξάνουν το πελατολόγιο τους άρα και τον τζίρο των καρτών. Στο εξωτερικό οι επιχειρηματικές πιστωτικές κάρτες κυκλοφορούν αρκετά χρόνια και φυσιολογικά είναι αρκετά διαδεδομένες. Η πρώτη εμφάνιση τέτοιας πιστωτικής κάρτας στην Ελλάδα έγινε το 1995 από την Εθνική Ασφαλιστική σε συνεργασία με την Mastercard, παρέχοντας ταυτόχρονα και ασφαλιστική κάλυψη στους κατόχους της. Στις μέρες μας κυκλοφορούν αρκετές τέτοιες κάρτες σε συνεργασία με πολυκαταστήματα, ασφαλιστικές εταιρείες, ποδοσφαιρικές ομάδες και άλλες επιχειρήσεις ή οργανισμούς, προσφέροντας έτσι αρκετές διευκολύνσεις. Αξιοπρόσεχτο είναι το γεγονός ότι στην παραγωγή νέων δανείων από τις τράπεζες η συμμετοχή των δικτύων λιανικών πωλήσεων ενισχύεται σε σημαντικό βαθμό με αποτέλεσμα το ένα τρίτο των καταναλωτικών δανείων που δίνουν οι τράπεζες να πωλείται μέσα από τις εμπορικές αλυσίδες καταστημάτων. Αντιλαμβανόμαστε λοιπόν, ότι τα περιθώρια για ανάπτυξη αυτής της αγοράς είναι μεγάλα, αν λάβουμε και υπόψη πως επεκτείνεται με ρυθμό τρεις φορές περίπου πιο γρήγορα από το μέσο όρο στην Ευρώπη. Σε αυτό το σημείο αξίζει επίσης να σημειωθεί ότι οι εξελίξεις στην ενιαία αγορά πληρωμών της ευρωζώνης ανοίγει το δίαυλο στις ίδιες τις επιχειρήσεις να εκδίδουν κάρτες πληρωμών χωρίς τη μεσολάβηση των τραπεζών, κάτι που σημαίνει ότι μπορούν να προχωρήσουν σε αυτό το βήμα από εταιρείες πετρελαίου μέχρι και σουπερμάρκετ.

2.5 ΕΞΟΡΥΞΗ ΓΝΩΣΗΣ ΚΑΙ ΕΠΑΓΓΕΛΜΑΤΙΚΕΣ ΧΡΗΜΑΤΟΔΟΤΗΣΕΙΣ

2.5.1 ΕΙΔΗ ΠΙΣΤΩΣΗΣ

Υπολογίζεται ότι τουλάχιστον μια στις τρεις μικρομεσαίες επιχειρήσεις καταφεύγει σήμερα στο τραπεζικό σύστημα για χρηματοδότηση, αν και έχουν αναπτυχθεί σύγχρονοι χρηματοδοτικοί μηχανισμοί για νέες επιχειρήσεις τα τελευταία χρόνια.

Οι τράπεζες δίνουν δύο ειδών δάνεια προς τις επιχειρήσεις:

- ✓ **Τα μακροπρόθεσμα:** Στα μακροπρόθεσμα περιλαμβάνονται τα δάνεια επαγγελματικού εξοπλισμού, αυτά δηλαδή που χρειάζονται για να καλυφθούν οι

ανάγκες εξοπλισμού μιας επιχείρησης και τα δάνεια εγκατάστασης, τα οποία καλύπτουν την ανάγκη απόκτησης επαγγελματικής στέγης.

- ✓ **Τα κεφαλαίου κίνησης:** Τα συγκεκριμένα δάνεια στοχεύουν στη βελτίωση της ρευστότητας της επιχείρησης και είναι μικρής διάρκειας.

Εκτός των άλλων πρέπει να τονίσουμε ένα σημαντικό χαρακτηριστικό των τραπεζικών δανείων το οποίο είναι η λογική των εμπραγμάτων ασφαλειών. Αυτό σημαίνει ότι οι τράπεζες δανείζουν μόνο σε όσους έχουν κάποιο περιουσιακό στοιχείο, το οποίο θα χρησιμοποιηθεί ως εγγύηση για την εξόφληση ολόκληρου ή μέρους του δανείου στην περίπτωση που ο δανειολήπτης δεν ανταποκριθεί στις υποχρεώσεις του. Σε περίπτωση λοιπόν που ένας επιχειρηματίας θέλει να ξεκινήσει μια προσπάθεια δίχως να έχει προσωπική περιουσία είναι αναγκαία η ύπαρξη ενός τρίτου προσώπου που να εγγυηθεί την δική του περιουσία. Συνειδητοποιούμε ότι οι διαφορετικές κατηγορίες πίστωσης είναι αρκετές και δύσκολα κατηγοριοποιούνται σε μικρότερες ομάδες δεδομένου της πολυπλοκότητας τους. Η διαφορετικότητα των τραπεζικών προϊόντων μπορεί να οφείλεται είτε στο χρόνο διάρκειας και στην σταθερότητα των δόσεων, είτε στο επιτόκιο και στα ενέχυρα ανταλλάγματα, είτε στον σκοπό του δανείου. Ωστόσο είναι απαραίτητο να αναφέρουμε ότι ο ενδιαφερόμενος δανειολήπτης είναι ορθό να εξετάσει τις δυνατότητες χρηματοδότησης του δανείου του με σταθερό ή κυμαινόμενο επιτόκιο, ανάλογα με την εξέλιξη του πληθωρισμού, των επιτοκίων και των υπόλοιπων οικονομικών μεγεθών. Πάντως το επιτόκιο αποπληρωμής του μακροπρόθεσμου δανείου είναι στις περισσότερες περιπτώσεις χαμηλότερο, από το αντίστοιχο επιτόκιο αποπληρωμής του δανείου κεφαλαίου κίνησης, εξαιτίας της μεγαλύτερης διάρκειας αποπληρωμής του.

2.5.2 ΔΑΝΕΙΣΜΟΣ ΚΑΙ ΑΞΙΟΛΟΓΗΣΗ ΤΗΣ ΠΙΣΤΟΛΗΠΤΙΚΗΣ ΙΚΑΝΟΤΗΤΑΣ ΤΩΝ ΕΠΙΧΕΙΡΗΣΕΩΝ

Οι τράπεζες από το 2008 εφαρμόζουν το αναθεωρημένο πλαίσιο κεφαλαιακής επάρκειας, το οποίο τις υποχρεώνει να πραγματοποιούν πιο λεπτομερή αξιολόγηση στην πιστοληπτική ικανότητα των επιχειρήσεων, για τη χορήγηση δανείων σε αυτές. Εξαιτίας αυτού του πλαισίου δημιουργήθηκε η ανάγκη για αυξημένες οργανωτικές προσαρμογές των επιχειρήσεων. Ταυτόχρονα ο σημαντικότερος σκοπός του προγράμματος αυτού, είναι η διευκόλυνση στην πρόσβαση, καθώς και η ενίσχυση της μικρομεσαίας επιχείρησης από την τραπεζική χρηματοδότηση με πολύ καλυτέρους όρους, σύμφωνα βεβαίως με τις απαιτήσεις του εκάστοτε οικονομικού περιβάλλοντος και ειδικά ως προς την αξιολόγηση της πιστοληπτικής ικανότητας της.

Επομένως, ως στόχοι του πλαισίου που προαναφέραμε μπορεί να θεωρηθούν οι εξής:

- o Ο επιχειρηματίας με μεθοδικότητα να παρακολουθεί και να καταλαβαίνει τις παραμέτρους που επηρεάζουν την πιστοληπτική ικανότητα της επιχείρησης του και να μπορεί να γίνεται αποτελεσματικότερος ως προς αυτή.

- o Ο δανειολήπτης να μπορεί να παρέχει τεκμηριωμένη και έγκαιρη πληροφόρηση κατά την διάρκεια χορήγησης του δανείου.

- Ο συνεπής πιστούχος να μπορεί να επιβραβευθεί με μειωμένες κεφαλαιακές απαιτήσεις από την τράπεζα στην εξυπηρέτηση των δανειακών του υποχρεώσεων.

2.5.3 ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΜΕΡΙΣΜΑ

Το μέρισμα είναι αυτό που αποκομίζει ένας επενδυτής για την επένδυση του σε μια επιχείρηση μέσω της αγοράς των μετοχών της. Η πολιτική μερισμάτων απασχολεί τα Διοικητικά Συμβούλια όλων των επιχειρήσεων και λαμβάνεται υπόψη προκειμένου κάθε επιχειρηματική οντότητα να λάβει ορθές αποφάσεις για τη μελλοντική πορεία της. Η μερισματική πολιτική αποτελείται από δύο κύριες αποφάσεις, εάν θα πληρωθεί μέρισμα στους μετόχους και εάν ναι πόσο; Η πρόβλεψη και μοντελοποίηση των δύο αυτών αποφάσεων έχει λάβει ένα σημαντικό ενδιαφέρον στην βιβλιογραφία. Ωστόσο, οι μέθοδοι που χρησιμοποιούνται για την εξέταση των προαναφερθέντων αποφάσεων περιορίζονται στην λογιστική παλινδρόμηση χωρίς καμία έρευνα να εφαρμόζει μεθόδους εξόρυξης δεδομένων.

Αυτές οι μέθοδοι έχουν αποδείξει την ανωτερότητα τους σε άλλα επιχειρησιακά πεδία όπως, μάρκετινγκ, παραγωγή, λογιστική, και ελεγκτική. Στην χρηματοοικονομική διοίκηση τη μερίδα του λέοντος κατέχει η πρόβλεψη της πτώχευσης μιας επιχείρησης αλλά στην μερισματική πολιτική δεν υπάρχει καμία τέτοια μέθοδος που να έχει εφαρμοστεί. Η επιτυχής πρόβλεψη της μερισματικής πολιτικής παρέχει πλεονεκτήματα σε όλα τα σχετιζόμενα μέρη καθώς μπορούν να διαχειριστούν, να συμβουλεύσουν και να παρακολουθήσουν την μερισματική πολιτική με έναν πιο αποτελεσματικό τρόπο.

2.6 ΕΦΑΡΜΟΓΕΣ ΕΞΟΡΥΞΗΣ ΓΝΩΣΗΣ ΣΕ ΧΡΗΜΑΤΟΟΙΚΟΝΟΜΙΚΟ ΠΕΡΙΒΑΛΛΟΝ

Σε αυτό το σημείο θα επικεντρωθούμε στο κατά πόσο οι τεχνικές εξόρυξης δεδομένων μπορούν να εφαρμοστούν πάνω σε οικονομικής φύσης προβλήματα και ζητήματα. Θα παρουσιάσουμε στοιχεία και πληροφορίες σχετικά με το πώς δημιουργείται ένα μοντέλο πρόβλεψης και εκτίμησης καθώς και τα στάδια προεργασίας που απαιτούνται.

2.6.1 ΔΗΜΙΟΥΡΓΙΑ ΕΝΟΣ ΜΟΝΤΕΛΟΥ ΠΡΟΒΛΕΨΗΣ

Αρχικά πρέπει να αναλογιστούμε πως μπορούμε να κατασκευάσουμε διάφορα υπολογιστικά μοντέλα τα οποία θα παίρνουν σαν είσοδο οικονομικά δεδομένα χρησιμοποιώντας τεχνικές εξόρυξης δεδομένων και θα επιστρέφουν σαν έξοδο μία εκτίμηση πρόβλεψης για το ανάλογο μέγεθος που εξετάζουμε. Η τεχνική εξόρυξης δεδομένων όπως προαναφέραμε απορρέει από τα πεδία της μηχανικής μάθησης και τεχνολογιών που αφορούν ανάλυση βάσεων δεδομένων (Database technologies). Ο σκοπός της μηχανικής μάθησης είναι να κατασκευαστούν υπολογιστικά προγράμματα τα οποία αυτόματα θα βελτιώνονται όσο αποκτούν πείρα και να παράγουν χρήσιμες εκτιμήσεις. Γενικά οι τεχνικές εξόρυξης δεδομένων όπως είναι τα δέντρα αποφάσεων και τα νευρωνικά δίκτυα που αναφέραμε στο προηγούμενο κεφάλαιο, μπορούν να αποδειχθούν ιδιαίτερα χρήσιμες στον κάθε ενδιαφερόμενο οικονομικό αναλυτή.

Παρακάτω παραθέτουμε ένα παράδειγμα σε μορφή πίνακα, για το που θα μπορούσε να εφαρμοστεί ένα μοντέλο εξόρυξης δεδομένων.

Εξόρυξη δεδομένων και εμπόριο	
Χρήστες	Σκοπός, εφαρμογή και αποτέλεσμα συστήματος
Έμποροι	• Προειδοποίηση για αλλαγή εμπορευμάτων • Εύρεση κανόνων και συσχετίσεων στην αγορά • Εύρεση σχέσεων μεταξύ των αγορών (κανόνες που αναφέρονται στο πως αλληλοεπηρεάζονται)
Manager	• Βελτίωση των εμπορικών διαδικασιών • Δημιουργία συνεχής ομαλότητας στις εμπορικές λειτουργίες

Παραδοσιακά Knowledge - based expert systems για εμπόρους	
Στόχος	• Έγκυρη προειδοποίηση για αλλαγή εμπορευμάτων • Εύρεση κανόνων στην αγορά • Εύρεση αλληλεπιδράσεων μεταξύ των αγορών
Δυσκολίες	• Επαλήθευση για την εγκυρότητα τους, εμπισοσύνη
Αποτέλεσμα	• Οι κανόνες που ισχύουν σε αρκετές διαφορετικές αγορές είναι σχετικά λίγοι

Γενικά οι τεχνικές της μεθόδου εξόρυξης από δεδομένα τείνουν να απαιτούν περισσότερα ιστορικά δεδομένα από τα κοινά μοντέλα και ειδικά στην περίπτωση των νευρωνικών δικτύων, είναι δύσκολο να ερμηνευθούν. Οι παράμετροι που λαμβάνονται υπόψη σε ένα σύστημα μοντέλο εξόρυξης δεδομένων έχουν ως εξής:

- **Data set:** Σε αυτή τη περίπτωση έχουμε δύο επιλογές. Είτε να χρησιμοποιήσουμε την χρονοσειρά που έχουμε στην διάθεσή μας, είτε όλες τις μεταβλητές που επηρεάζουν την χρονοσειρά. Οι τεχνικές εξόρυξης δεδομένων ακολουθούν μία θεμελιώδη προσέγγιση ανάλυσης (fundamental analysis approach) η οποία έχει να κάνει με τους σημαντικότερους παράγοντες επιρροής.

- **Data types:** Συνήθως οι μέθοδοι εξόρυξης δεδομένων ακολουθούν μία χαρακτηριστική προσέγγιση attributevalue approach (attribute-based approach). Αυτή η μέθοδος χρησιμοποιεί και περιλαμβάνει πολλές στατιστικές μεθόδους καθώς και νευρωνικά δίκτυα.

- **Mathematical algorithm method, model:** Αποτελεί μία ποικιλία από στατιστικές και νευρωνικές μεθόδους καθώς και μεθόδους που έχουν να κάνουν με την λογική. Υπάρχουν πολλά νευρωνικά δίκτυα τα οποία βασίζονται σε διαφορετικούς αλγόριθμους και μεθοδολογίες.

Στο προηγούμενο κεφάλαιο είχαμε κάνει μια αναφορά στα τρία βασικά βήματα που καθορίζουν την δημιουργία ενός υπολογιστικού μοντέλου πρόβλεψης. Παρακάτω θα αναφερθούμε πιο συγκεκριμένα στις πληροφορίες για τα στοιχεία που περιέχουν τα στάδια αυτά, καθώς και για το πώς επενεργούν στο όλο σύστημα πρόβλεψης.

2.6.2 ΠΡΟ-ΕΠΕΞΕΡΓΑΣΙΑ ΔΕΔΟΜΕΝΩΝ

Προτού τα διάφορα δεδομένα δοθούν σαν είσοδο σε έναν αλγόριθμο, πρέπει να συλλεχθούν, να ερευνηθούν και να καθαριστούν. Ακόμα και το καλύτερο σύστημα πρόβλεψης θα αποτύχει στην σωστή εξαγωγή συμπερασμάτων αν τα δεδομένα που πραγματεύεται είναι κακής ποιότητας. Επίσης είναι σκόπιμο να εξεταστούν ποια δεδομένα κρίνονται σαν καταλληλότερα για αξιοποίηση ώστε να έχουμε το καλύτερο δυνατό αποτέλεσμα. Πολλές φορές κάποια τιμή απουσιάζει από την βάση δεδομένων που διαθέτουμε για εκμετάλλευση. Είναι προτιμότερο να αντικατασταθεί με μία άλλη τιμή παρά να διαγραφεί. Σε περίπτωση όπου έχουμε μεγάλο αριθμό τέτοιων εγγραφών τότε είναι καλό να τις αντικαταστήσουμε με εκείνη την τιμή που θα επηρεάσει λιγότερο την εγκυρότητα της πρόβλεψής μας. Επίσης μπορούν να χρησιμοποιηθούν κάποιοι δείκτες (indicators) των οποίων η χρησιμότητα είναι πολύ μεγάλη καθώς μειώνουν τον θόρυβο (από την στιγμή που εκφράζουν έναν μέσο όρο) και παρέχουν όψεις από δεδομένα τα οποία είναι κατάλληλα για επεξεργασία. Πολλές φορές επιβάλλεται η αναγνώριση κάποιων χαρακτηριστικών από το μοντέλο όσον αφορά τα δεδομένα και σε αυτήν την περίπτωση χρησιμοποιούνται τεχνικές όπως Ανάλυση κύριων συνιστωσών (Principal component analysis), Ανάλυση ευαισθησίας (Sensitivity analysis) και κάποιες άλλες τεχνικές όπως είναι οι Ευρετικές (Heuristic).

2.6.3 ΕΠΙΛΟΓΗ ΚΑΤΑΛΛΗΛΩΝ ΑΛΓΟΡΙΘΜΩΝ ΚΑΙ ΕΠΙΛΟΓΗ ΜΟΝΤΕΛΟΥ

Όσον αφορά τις μεθόδους και τους αλγορίθμους που χρησιμοποιούνται για την πρόβλεψη χρηματοοικονομικών μεγεθών υπάρχει πληθώρα τέτοιων τεχνικών οι σημαντικότερες εκ των οποίων είναι οι γραμμικές μέθοδοι και τα νευρωνικά δίκτυα. Μελετώντας αυτές τις τεχνικές το συμπέρασμα που βγαίνει είναι ότι οι καταλληλότερες είναι οι υβριδικές, δηλαδή συνδυασμός πολλών τεχνικών για την αξιοποίηση των δυνατών σημείων της κάθε μίας. Μία άλλη προσέγγιση που σχετίζεται, με την μέθοδο εξόρυξης δεδομένων, είναι να υιοθετηθεί ένα μοντέλο το οποίο είναι ευέλικτο στο ότι μπορεί να συνδυάσει έναν μεγάλο αριθμό από συναρτήσεις με μεγάλη ακρίβεια. Τέτοια μοντέλα είναι μη παραμετρικά καθώς δεν χρειάζεται να υπάρχει άμεση σχέση μεταξύ των τιμών των παραμέτρων ενός μοντέλου με δεδομένα.

Τα πλεονεκτήματα αυτού του είδους των μοντέλων είναι τα παρακάτω:

1. Η δυνατότητα που παρέχουν στο να μοντελοποιούν υψηλής πολυπλοκότητας συναρτήσεις

2. Η δυνατότητα να χρησιμοποιούν έναν υψηλό αριθμό μεταβλητών στο μοντέλο, και παρ' όλα αυτά να περιέχουν και άλλα δεδομένα όπως θεμελιώδεις και τεχνικούς παράγοντες.

Σαν μειονέκτημα για αυτά τα μοντέλα μπορεί να θεωρηθεί ότι δεν μπορούν να ερμηνευθούν εύκολα.

2.6.4 ΚΑΘΑΡΙΣΜΟΣ ΟΙΚΟΝΟΜΙΚΩΝ ΔΕΔΟΜΕΝΩΝ

Συνεχώς καινούργιες τεχνικές και μέθοδοι έχουν δημιουργηθεί ώστε να είμαστε σε θέση να αναλύουμε οικονομικά δεδομένα και να είμαστε σε θέση να λαμβάνουμε έπειτα τις όποιες αποφάσεις. Σε πολλές περιπτώσεις όμως τα δεδομένα που χρησιμοποιούν αυτές οι μέθοδοι προς ανάλυση είναι γεμάτα από λάθη. Αυτά ονομάζονται βρώμικα δεδομένα (dirty data). Επίσης, είναι γεγονός ότι οι περισσότερο εξελιγμένες και συνάμα πιο πολύπλοκες τεχνικές δείχνουν να επηρεάζονται από τέτοιου είδους ατελή δεδομένα και να μας δημιουργούν την ανάγκη να τα επεξεργαστούμε για να καθαριστούν και να είναι έτοιμα για ανάλυση.

Υπάρχουν οι εξής περιπτώσεις που μπορούμε να συναντήσουμε αντιμετωπίζοντας ατελή δεδομένα:

- Να μην έχουν τοποθετηθεί τιμές κατά την διάρκεια των εγγραφών. Αυτό γίνεται πολύ εύκολά κατανοητό καθώς αποτελεί την περίπτωση να έχουμε κενά κελιά τιμών δηλαδή έλλειψη δεδομένων.

- Να αναφέρονται σε τιμές που είναι αδύνατο να ισχύουν, οι λεγόμενες impossible values, όπως σε περιπτώσεις που αναμένουμε θετική τιμή και υπάρχει αρνητική. Στην

περίπτωση κατά την οποία οι σωστές τιμές δεν μπορούν να καταχωρηθούν, τότε θα πρέπει να καταφύγουμε στην παρατήρηση τιμών που απουσιάζουν και λείπουν.

- Να αναφέρονται σε τιμές οι οποίες αποτελούν ένα πιο πολύπλοκο λάθος, οι λεγόμενες, inconsistent values, και αυτό συμβαίνει όταν πολλές τιμές μαζί σπάνε έναν κανόνα. Για παράδειγμα εάν κάποιες τιμές που αποτελούν συστατικά ενός κανόνα δεν έχουν ως αποτέλεσμα την εξαγωγή του κανόνα, ένας τρόπος για να αντιμετωπιστεί αυτή η κατάσταση είναι να θεωρήσουμε εμείς ποιες από τις τιμές είναι οι κατάλληλες.

Για παράδειγμα απίθανη τιμή είναι εκείνη που σε μία ιεραρχία μας ξενίζει όπως στην 2,3,5,7,10,2000. Η τιμή 2000 μπορεί να είναι σωστή αλλά είναι απίθανη να ισχύει εδώ. Πιθανόν να έγινε λάθος στην καταχώρησή της και το σύστημα θα πρέπει να την αναγνωρίζει και να την θεωρεί άκυρη. Τι γίνεται όμως στην περίπτωση όπου αντί 2000 θα είχαμε 200. Μήπως θα πρέπει να ερευνηθεί η τιμή λίγο παραπάνω καθώς η απόκλιση δεν είναι και τόσο μεγάλη;

Όπως παρατηρούμε, αρκετών ειδών περιπτώσεις λαθών μπορούμε να υπάρξουν αλλά η αντιμετώπισή τους εκτός ότι μπορεί να μην είναι εύκολη είναι συνάμα και χρονοβόρα. Αρχικά θα αναφερθούμε στην περίπτωση που έχουμε ελλιπή δεδομένα (missing data). Στην περίπτωση αυτή το πρώτο που κάνουμε είναι να διαπιστώσουμε την παρατήρηση η οποία είναι περισσότερο κοντά στην τιμή που λείπει. Αυτό αποκαλείται ως δότης (donor). Εάν υπάρχουν περισσότερα πιθανά από ένα donor τότε μπορούμε να επιλέξουμε ένα από αυτά στην τύχη. Στην περίπτωση αυτή προσπαθούμε να ταιριάξουμε την κενή τιμή με το κατάλληλο donor.

Η δεύτερη προσέγγιση είναι να χρησιμοποιήσουμε κάποιο μοντέλο για να βρούμε την τιμή που απουσιάζει. Η βασική και θεμελιώδης προσέγγιση είναι η ακόλουθη:

- Δημιουργία ενός μοντέλου
- Χρήση του μοντέλου για ανακάλυψης της τιμή που λείπει

Όπως συμβαίνει συνήθως, το βασικό θέμα είναι να επιλέξουμε το ιδανικότερο μοντέλο. Όσο ιδανικότερο το μοντέλο, τόσο καλύτερο το αποτέλεσμα. Το μοντέλο θα περιέχει όλα τα δεδομένα και τις αντίστοιχες χρονοσειρές και επίσης θα πρέπει να είναι όσο το δυνατόν απλό στην χρήση του και στην δομή του. Για να δημιουργήσουμε ένα μοντέλο θα πρέπει να συμπληρώσουμε δεδομένα. Θα πρέπει όσον το δυνατό να έχουμε περισσότερες μη κενές τιμές στα κελιά. Στην περίπτωση που δε συμβαίνει αυτό, μπορούμε να το αντιμετωπίσουμε με το να δοκιμάζουμε και να προβλέπουμε διάφορες τιμές. Στην περίπτωση που τα αποτελέσματα έχουν την βέλτιστη επιτυχία και εγκυρότητα έχουμε τον κατάλληλο αλγόριθμο.

2.6.5 ΑΠΟΤΙΜΗΣΗ ΤΟΥ ΣΥΣΤΗΜΑΤΟΣ ΠΟΥ ΔΗΜΙΟΥΡΓΗΣΑΜΕ

Ένα από τα βασικότερα στάδια κατά την δημιουργία και την υλοποίηση ενός μοντέλου πρόβλεψης είναι η αποτίμηση του μοντέλου, δηλαδή πόσο καλά και ικανοποιητικά είναι τα αποτελέσματα ή οι πληροφορίες τις οποίες αποδίδει. Αυτό συμβαίνει διότι η αποτίμηση μας δείχνει που εστιάζεται το ενδιαφέρον του μοντέλου, εάν το μοντέλο μας είναι ιδανικό, θα μας παρουσιάσει τα οφέλη τα οποία σε άλλη περίπτωση δε θα μπορούσαν να αποκομιστούν και τέλος αιτιολογεί και αποδεικνύει πως εξήχθησαν τα οποιαδήποτε συμπεράσματα και πληροφορίες.

Στις προβλέψεις που αφορούν χρηματοοικονομικά δεδομένα ο στόχος της πρόβλεψης είναι η κερδοφορία ενώ οι αλγόριθμοι οι οποίοι μπορεί να χρησιμοποιηθούν ίσως έχουν άλλο στόχο. Για τον λόγο αυτό πρέπει οι αλγόριθμοι και οι μέθοδοι οι οποίες θα χρησιμοποιηθούν να εξετάζονται προηγουμένως για το πόσο είναι κατάλληλες σε τέτοιου είδους δεδομένα. Υπάρχουν ορισμένες στρατηγικές οι οποίες έχουν αναπτυχθεί για αυτόν ακριβώς τον λόγο όπως:

i. Ακρίβεια (Accuracy), η οποία δηλώνει ένα ποσοστό σωστών αποτελεσμάτων

ii. Τετραγωνικό σφάλμα (Square error), το άθροισμα των διακυμάνσεων των αποτελεσμάτων

iii. Αξιοπιστία (Reliability), η αξιοπιστία του μοντέλου στην πρόβλεψη

2.7 ΔΙΑΧΕΙΡΙΣΗ ΠΕΛΑΤΕΙΑΚΩΝ ΣΧΕΣΕΩΝ (CRM) ΚΑΙ ΕΞΟΡΥΞΗ ΓΝΩΣΗΣ

Τα συστήματα διαχείρισης των σχέσεων μιας εταιρείας με τους πελάτες της (Customer Relationship Management Systems, CRMS), άρχισαν να πρωτοεμφανίζονται στην αγορά στις αρχές τις δεκαετίας του ’90 και κατάφεραν να αποκτήσουν ένα σημαντικό ρόλο στη ζωή των εταιρειών στα τέλη αυτής. Για πολλά χρόνια, οι επιχειρήσεις είχαν επικεντρώσει τις προσπάθειες τους στην περικοπή των λειτουργικών εξόδων και στην αύξηση της απόδοσης στα πλαίσια του ίδιου του οργανισμού. Κατά καιρούς προσπάθησαν να απλοποιήσουν και να αυξήσουν την αποδοτικότητα των εσωτερικών διαδικασιών, αυτοματοποιώντας κάποια από τα στοιχεία των καλούμενων back office λειτουργιών, όπως είναι το κατασκευαστικό κομμάτι, η διαχείριση αποθήκης και τα οικονομικά.

Η εξόρυξη δεδομένων μπορεί να βοηθήσει τις επιχειρήσεις να ανταπεξέλθουν σε περιόδους αιχμής της κατανάλωσης, στη διακίνηση των εμπορευμάτων, αλλά και στη μείωση των ζημιών που οφείλονται σε εσωτερικές απάτες (internal fraud), οι οποίες υπολογίζεται ότι είναι περίπου το 40% έως 50 % των αποθεμάτων των επιχειρήσεων λιανικής πώλησης. Η εξόρυξη δεδομένων μπορεί επίσης να βοηθήσει στην επίλυση ασυνήθιστων μοτίβων που αφορούν επιστροφές εμπορευμάτων, εκπτώσεις, υπερισχύουσες τιμές, πιστωτικές κάρτες, κάρτες καταστημάτων, χρεωστικές κάρτες, εκπτώσεις που παρέχονται στο προσωπικό, πλεονάζοντα εμπορεύματα και ελλείψεις που οφείλονται στα αποθέματα που αναφέρονται ως

κατεστραμμένα ή ελαττωματικά, καθιστώντας με αυτόν τον τρόπο την ανίχνευση της λιανικής απάτης πολύ πιο εύκολη, ακριβής, έγκαιρη και οικονομική.

Όπως φαίνεται λοιπόν και στο παρακάτω σχήμα, η χρήση της εξόρυξης δεδομένων πάνω σε συστήματα διαχείρισης των σχέσεων μιας εταιρείας με τους πελάτες της, μπορεί να απόφέρει ανταγωνιστικά πλεονεκτήματα έναντι άλλων επιχειρήσεων ή οργανισμών.

Στη συνέχεια, γίνεται η ταξινόμηση των τεχνικών εξόρυξης δεδομένων που εφαρμόζονται στη διαχείριση πελατειακών σχέσεων, αναφέρονται τα βήματα για την προετοιμασία των δεδομένων, προκειμένου αυτά να εισέλθουν στη διαδικασία της εξόρυξης, και αναλύονται τα στάδια για την αποτελεσματική εφαρμογή της εξόρυξης δεδομένων στο CRM. Στη συνέχεια, παρουσιάζονται το μοντέλο τμηματοποίησης πελατών και το σύστημα δημιουργίας προφίλ πελατών χρησιμοποιώντας τα εργαλεία της εξόρυξης δεδομένων. Τέλος, γίνεται αναφορά στο Marketing Data Intelligence, το οποίο είναι το αποτέλεσμα της εφαρμογής της εξόρυξης δεδομένων στον κύκλο ζωής του πελάτη.

2.7.1 CUSTOMER RELATIONSHIP MANAGEMENT (CRM)

Τα τελευταία χρόνια έχουν γίνει σημαντικές βελτιώσεις, όσον αφορά την ενσωμάτωση της εξόρυξης δεδομένων στη διαδικασία του CRM. Η τάση αυτή αναμένεται να συνεχιστεί, με αποτέλεσμα οι εφαρμογές του CRM να έχουν όλο και περισσότερες δραστηριότητες μάρκετινγκ βασιζόμενες στα αποτελέσματα της εξόρυξης δεδομένων (Thearling, 2010). Ενώ οι μελέτες για την εξόρυξη δεδομένων έχουν επικεντρωθεί κυρίως στις τεχνικές, οι μελέτες για τις πελατειακές σχέσεις έχουν επικεντρωθεί στην αλληλεπίδραση του πελάτη και των στρατηγικών της εταιρείας. Η σωστή διαχείριση πελατειακών σχέσεων μπορεί να επιτευχθεί μόνο μέσω της ενοποίησης της διαδικασίας ανακάλυψης γνώσης με τη διαχείριση και χρήση της γνώσης στις στρατηγικές μάρκετινγκ. Αυτό βοηθάει την επιχείρηση στην αντιμετώπιση των αναγκών των πελατών, με βάση τις πληροφορίες που γνωρίζει για τον κάθε πελάτη ξεχωριστά, παρά με τη χρήση μίας μαζικής γενίκευσης των χαρακτηριστικών των πελατών (Shaw et al., 2001).

2.7.2 ΤΑΞΙΝΟΜΗΣΗ ΤΩΝ ΤΕΧΝΙΚΩΝ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ ΣΤΙΣ ΤΕΣΣΕΡΙΣ ΔΙΑΣΤΑΣΕΙΣ ΤΟΥ CRM

Η ανάλυση και η κατανόηση των συμπεριφορών και των χαρακτηριστικών των πελατών αποτελούν τα θεμέλια της ανάπτυξης μιας ανταγωνιστικής στρατηγικής CRM με σκοπό την

απόκτηση και τη διατήρηση των δυνητικών πελατών και τη μεγιστοποίηση της αξίας των πελατών. Τα εργαλεία εξόρυξης δεδομένων, τα οποία είναι κατάλληλα στο να εξάγουν και να εντοπίζουν χρήσιμες πληροφορίες και γνώσεις από την τεράστια βάση δεδομένων των πελατών, προσφέρουν μεγάλη υποστήριξη στη λήψη διαφορετικών αποφάσεων CRM (Ngai et al,, 2009). Πολύ συχνά, απαιτείται να γίνει συνδυασμός των μοντέλων εξόρυξης δεδομένων για την υποστήριξη ή την πρόβλεψη των επιπτώσεων μίας στρατηγικής CRM. Σε μία τέτοια περίπτωση, η ταξινόμηση των μοντέλων εξόρυξης δεδομένων θα βασίζεται στα σημαντικότερα σημεία της στρατηγικής CRM (Ngai et al. 2009). Για παράδειγμα, δεδομένου ότι οι σχέσεις μεταξύ των προϊόντων είναι το κύριο μέλημα της επιχείρησης, στην περίπτωση των προγραμμάτων σταυροειδών πωλήσεων, οι πελάτες μπορούν να κατηγοριοποιηθούν σε ομάδες προτού εφαρμοστεί σε κάθε ομάδα κάποιο μοντέλο συσχετίσεων. Σε τέτοιες περιπτώσεις, το πρόγραμμα σταυροειδών πωλήσεων θα πρέπει να ταξινομηθεί με βάση το μοντέλο συσχετίσεων. Στην περίπτωση του άμεσου μάρκετινγκ, ένα ορισμένο τμήμα των πελατών μπορεί να υποδιαιρεθεί σε ομάδες, έτσι ώστε να σχηματιστούν τα αρχικά τμήματα του μοντέλου τμηματοποίησης. Το πρόγραμμα του άμεσου μάρκετινγκ θα ταξινομηθεί με βάση την τμηματοποίηση, δεδομένου ότι το κύριο μέλημα εδώ είναι η πρόβλεψη της συμπεριφοράς των πελατών.

2.7.3 ΠΡΟΕΤΟΙΜΑΣΙΑ ΤΩΝ ΔΕΔΟΜΕΝΩΝ ΓΙΑ ΤΗ ΔΙΑΔΙΚΑΣΙΑ ΤΗΣ ΕΞΟΡΥΞΗΣ ΓΝΩΣΗΣ

2.7.3.1 ΠΡΟΕΤΟΙΜΑΣΙΑ ΤΩΝ ΔΕΔΟΜΕΝΩΝ ΓΙΑ ΟΛΑ ΤΑ ΕΡΓΑΛΕΙΑ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ

Πολύ συχνά, απαιτείται να γίνει προετοιμασία των δεδομένων, προκειμένου να εισέλθουν στην διαδικασία της εξόρυξης δεδομένων. Μάλιστα, μπορεί να δαπανηθεί περισσότερος χρόνος κατά την προετοιμασία των δεδομένων, παρά κατά τη διαδικασία της εξόρυξης δεδομένων. Τα κυριότερα βήματα, προκειμένου τα δεδομένα να αποκτήσουν την απαιτούμενη μορφή, είναι τα εξής (Kimball, 1997):

Η διόρθωση μη συνεπούς μορφής δεδομένων και η διόρθωση μη συμβατής κωδικοποίησης των δεδομένων, συντομογραφιών και σημείων στίξης.

- ✓ Η αφαίρεση ανεπιθύμητων ή περιττών πεδίων. Τα δεδομένα μπορεί να περιέχουν πεδία άνευ σημασίας για την ανάλυση που θέλουμε να κάνουμε. Τα εργαλεία της εξόρυξης δεδομένων ενδέχεται να ερμηνεύσουν αυτά τα πεδία ως μετρήσεις ή μεγέθη, ειδικά αν είναι αριθμητικά, και μπορεί να κάνουν κύκλους προσπαθώντας να βρουν συγκεκριμένα μοτίβα με αυτά τα πεδία, ή προσπαθώντας να συσχετίσουν αυτά τα πεδία με πραγματικά δεδομένα.

- ✓ Η ερμηνεία των κωδικών σε κείμενο. Η κλασική μορφή καθαρισμού των δεδομένων περιλαμβάνει τη βελτίωση ή την αντικατάσταση αινιγματικών κωδικών με ισοδύναμα κειμένου, γραμμένο με αναγνωρίσιμες λέξεις.

- ✓ Ο συνδυασμός των δεδομένων που προέρχονται από διάφορες πηγές. Όπως τα δεδομένα των πελατών σε μία κοινή βάση.

- ✓ Η εύρεση των πεδίων που έχουν χρησιμοποιηθεί για παραπάνω από έναν σκοπό. Ένας καλός τρόπος για να βρεθούν αυτά τα αρχεία είναι η καταμέτρηση, και ίσως και η δημιουργία μίας λίστας όλων των διαφορετικών τιμών-χρήσεων που υπάρχουν σε ένα πεδίο.

- ✓ Ο έλεγχος για τυχόν μη φυσιολογικά, εκτός ορίων ή αδύνατα στοιχεία. Κάποια μετρήσιμα στοιχεία μπορεί να είναι σωστά, αλλά εξαιρετικά ασυνήθιστα. Τέτοιου είδους δεδομένα, είναι προτιμότερο να μαρκαριστούν με μία ειδική σήμανση, έτσι ώστε να μπορούμε να τα συμπεριλάβουμε ή να τα εξαιρέσουμε από την ανάλυσή μας, ανάλογα με την περίπτωση.

- ✓ Ο έλεγχος για τυχόν τιμές που λείπουν, ή εάν αυτές έχουν αντικατασταθεί από κάποιον προεπιλεγμένο αριθμό.

- ✓ Η εφαρμογή ομοιόμορφης μεταχείρισης σε μηδενικές τιμές. Οι μηδενικές τιμές ενδέχεται να δυσκολέψουν τη λειτουργία του εργαλείου Εξόρυξης Δεδομένων. Σε πολλές περιπτώσεις, η μηδενική αξία αντιπροσωπεύεται από κάποια άλλη ιδιαίτερη τιμή. Πολλές φορές η τιμή -1 θεωρείται ότι αντιπροσωπεύει τις μηδενικές αξίες.

- ✓ Η ταξινόμηση μεμονωμένων αρχείων δεδομένων σύμφωνα με ένα από τα συγκεντρωτικά του μεγέθη. Σε ορισμένες περιπτώσεις, μπορεί να είναι επιθυμητός ο εντοπισμός της πώλησης ενός πολύ συγκεκριμένου προϊόντος, όπως ένα ρούχο σε ένα συγκεκριμένο χρώμα και μέγεθος, και από ένα συγκεκριμένο υλικό.

2.7.3.2 ΠΡΟΕΤΟΙΜΑΣΙΑ ΤΩΝ ΔΕΔΟΜΕΝΩΝ ΑΝΑΛΟΓΑ ΜΕ ΤΟ ΧΡΗΣΙΜΟΠΟΙΟΥΜΕΝΟ ΕΡΓΑΛΕΙΟ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ

Ανάλογα με το χρησιμοποιούμενο εργαλείο εξόρυξης δεδομένων, μπορεί να χρειαστεί να γίνουν κάποιες επιπλέον μετατροπές στα δεδομένα, πέρα των προηγουμένων, και αυτές είναι οι ακόλουθες:

- **Ο διαχωρισμός των πρωτογενών δεδομένων εισόδου σε τρεις ομάδες.** Η πρώτη ομάδα δεδομένων χρησιμοποιείται για την κατάρτιση του εργαλείου εξόρυξης δεδομένων. Ένα εργαλείο συσταδοποίησης (clustering tool), ένα εργαλείο νευρωνικών δικτύων (neural network tool) ή ένα εργαλείο δέντρου αποφάσεων (decision tree tool) απορροφά την πρώτη σειρά στοιχείων και ορίζει τις παραμέτρους, από τις οποίες μπορούν να γίνουν οι μελλοντικές ταξινομήσεις και οι προβλέψεις. Το δεύτερο σύνολο δεδομένων, που χρησιμοποιείται στη συνέχεια, ελέγχει αυτές τις παραμέτρους για να δει πόσο καλά αποδίδει το μοντέλο. Όταν το εργαλείο εξόρυξης δεδομένων έχει ρυθμιστεί σωστά στο πρώτο και στο δεύτερο σετ δεδομένων, εφαρμόζεται στη συνέχεια στην τρίτη σειρά αξιολόγησης των δεδομένων, όπου τα συμπλέγματα (clusters), οι ταξινομήσεις και οι προβλέψεις που προέρχονται από το εργαλείο είναι πλήρως αξιόπιστες και μπορούμε να τις χρησιμοποιήσουμε.

- **Η προσθήκη υπολογισμένων πεδίων ως εισροές ή ως στόχοι.** Για παράδειγμα, ένα υπολογιζόμενο πεδίο, όπως τα κέρδη ή η ικανοποίηση των πελατών, που αντιπροσωπεύει την αξία ενός συνόλου συναλλαγών των πελατών, μπορεί να χρειαστεί να τεθεί ως στόχος για να επιλέξει το εργαλείο εξόρυξης δεδομένων τους πιο κερδοφόρους πελάτες ή για να επιλέξει τη συμπεριφορά που θέλουμε να ενθαρρύνουμε.

- **Η διάταξη των συνεχών τιμών σε κλίμακες.** Μερικά εργαλεία εξόρυξης δεδομένων, όπως τα δέντρα αποφάσεων, ενθαρρύνουν τη διάταξη αυτή σε διακριτές κλίμακες.

- **Η εξομάλυνση των τιμών μεταξύ 0 και 1.** Τα εργαλεία νευρωνικών δικτύων συνήθως απαιτούν όλες οι αριθμητικές τιμές να αντιστοιχίζονται σε μια σειρά από το μηδέν έως το ένα.

- **Η μετατροπή των κειμένων σε αριθμητικές τιμές.** Μερικά εργαλεία εξόρυξης δεδομένων μπορούν να λειτουργήσουν μόνο με αριθμητικά δεδομένα εισόδου. Σε αυτές τις περιπτώσεις, οι διακριτές τιμές κείμενου θα πρέπει να αντικατασταθούν από ειδικούς κωδικούς, όπως για παράδειγμα η αντικατάσταση της περιοχής του κάθε πελάτη με τον αντίστοιχο ταχυδρομικό του κώδικα.

2.7.4 ΤΑ ΒΗΜΑΤΑ ΕΦΑΡΜΟΓΗΣ ΤΗΣ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ ΣΤΗ ΔΙΑΧΕΙΡΙΣΗ ΠΕΛΑΤΕΙΑΚΩΝ ΣΧΕΣΕΩΝ

Σύμφωνα με τον Edelstein (2000), τα βασικά βήματα της εξόρυξης δεδομένων για ένα αποτελεσματικό σύστημα διαχείρισης πελατών είναι τα ακόλουθα:

1. **Καθορισμός του προβλήματος της επιχείρησης (define business problem).** Κάθε εφαρμογή CRM έχει έναν ή περισσότερους επιχειρηματικούς στόχους για τους οποίους θα πρέπει να οικοδομηθεί το κατάλληλο μοντέλο. Ο αποτελεσματικός εντοπισμός του προβλήματος περιλαμβάνει και έναν τρόπο μέτρησης των αποτελεσμάτων του έργου του CRM.

2. **Δημιουργία βάσης δεδομένων μάρκετινγκ (build a marketing database).** Είναι απαραίτητη η οικοδόμηση μίας βάσης δεδομένων μάρκετινγκ, διότι οι επιχειρησιακές βάσεις δεδομένων συχνά δεν περιέχουν τα δεδομένα που απαιτούνται, με τη μορφή που απαιτούνται. Επίσης, εάν υπάρχουν ξεχωριστές βάσεις δεδομένων, όπως για παράδειγμα ξεχωριστή βάση δεδομένων για τους πελάτες, ξεχωριστή για τα προϊόντα και ξεχωριστή για τις συναλλαγές, θα πρέπει να γίνει μία ενσωμάτωση όλων αυτών σε μία ενιαία βάση μάρκετινγκ.

3. **Εξερεύνηση των δεδομένων (explore the data).** Απαραίτητη κρίνεται η σωστή κατανόηση των δεδομένων προκειμένου να οικοδομηθεί ένα αξιόπιστο μοντέλο προβλέψεων. Χρήσιμη θα μπορούσε να είναι η συγκέντρωση κάποιων αριθμητικών δεικτών, όπως για παράδειγμα μέσοι όροι και τυπικές αποκλίσεις, και η εξέταση της διανομής των δεδομένων. Σημαντική βοήθεια προσφέρουν οι γραφικές παραστάσεις και τα εργαλεία απεικόνισης. Η οπτικοποίηση των δεδομένων οδηγεί συχνά σε νέες ιδέες και κατά συνέπεια, στην επιτυχία.

4. **Προετοιμασία των δεδομένων για τη μοντελοποίηση ([prepare data for modeling).** Αρχικά επιλέγονται οι μεταβλητές πάνω στις οποίες θα κατασκευαστεί το μοντέλο και στη συνέχεια κατασκευάζονται οι νέες μεταβλητές πρόβλεψης, οι οποίες προέρχονται από τα ανεπεξέργαστα δεδομένα. Κατόπιν, επιλέγεται ένα υποσύνολο ή ένα δείγμα των δεδομένων, πάνω στα οποία θα κατασκευαστεί το μοντέλο. Το τελικό στάδιο αφορά τη μετατροπή των μεταβλητών σύμφωνα με τις απαιτήσεις του αλγορίθμου που θα επιλεχθεί για να κατασκευαστεί το μοντέλο.

5. **Κατασκευή του μοντέλου εξόρυξης δεδομένων (build model).** Η κατασκευή του μοντέλου αποτελεί μία επαναληπτική διαδικασία και θα πρέπει να διερευνηθούν εναλλακτικά μοντέλα για να βρεθεί το καταλληλότερο στην επίλυση του προβλήματος της επιχείρησης. Κατά τη διαδικασία αυτή ανίχνευσης του μοντέλου, πολλές φορές μπορεί να οδηγηθούμε σε κάποιο προηγούμενο βήμα και να κάνουμε κάποιες αλλαγές στα δεδομένα, ή ακόμα και να γίνει επαναπροσδιορισμός του αρχικού προβλήματος της επιχείρησης.

6. **Αξιολόγηση του μοντέλου (evaluate model).** Ίσως το πιο υπερεκτιμημένο μέτρο αξιολόγησης του μοντέλου είναι η ακρίβεια των αποτελεσμάτων. Ένα άλλο μέτρο που χρησιμοποιείται συχνά είναι η ανύψωση (lift), η οποία μετράει τη βελτίωση που επιτυγχάνεται με το μοντέλο πρόβλεψης. Ωστόσο, η μέθοδος αυτή δε λαμβάνει υπόψη τα κόστη και τα έσοδα, έτσι είναι συχνά προτιμότερο να εξετάσουμε τους δείκτες που αναφέρονται στα κέρδη ή την απόδοση της επένδυσης (ROI).

7. **Ανάπτυξη του μοντέλου και αποτελέσματα (deploy model and results).** Στην πραγματικότητα, ο τρόπος που ενσωματώνεται η εξόρυξη δεδομένων στην εφαρμογή του CRM καθορίζεται από το είδος της αλληλεπίδρασης της εταιρείας με τον πελάτη. Υπάρχουν δύο βασικοί τρόποι αλληλεπίδρασης: η εισερχόμενη, κατά την οποία οι πελάτες επικοινωνούν με την εταιρεία, και η εξερχόμενη, κατά την οποία η εταιρεία επικοινωνεί με τους πελάτες. Στις εισερχόμενες συναλλαγές, όπως για παράδειγμα μία τηλεφωνική παραγγελία ή παραγγελία μέσω internet, η εφαρμογή θα πρέπει να ανταποκριθεί σε πραγματικό χρόνο, κάτι το οποίο δε συμβαίνει στις εξερχόμενες συναλλαγές.

Στο επόμενο σχήμα παρουσιάζουμε τα βήματα για ένα αποτελεσματικό σύστημα διαχείρισης πελατών.

2.7.5 ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ ΣΤΟΝ ΚΥΚΛΟ ΖΩΗΣ ΤΟΥ ΠΕΛΑΤΗ

Για να είναι αποτελεσματικό ένα CRM σύστημα, θα πρέπει να συνδεθούν τα προϊόντα και οι στρατηγικές της εταιρείας με τους στόχους και τους πελάτες της. Ο όρος κύκλος ζωής του πελάτη (customer life cycle) αναφέρεται σε όλα τα στάδια της σχέσης μεταξύ της επιχείρησης και του πελάτη. Είναι απολύτως απαραίτητη η κατανόησή του όρου από την εταιρεία, διότι σχετίζεται άμεσα με την κερδοφορία της (Rygielski et al., 2002).

Σύμφωνα με τον Edelstein (2000), ο κύκλος ζωής του πελάτη αποτελείται από τρία στάδια, τα οποία είναι:

- **Απόκτηση πελατών:** Αποτελεί το πρώτο βήμα του CRM και η εξόρυξη δεδομένων μπορεί να συνεισφέρει στη βελτίωση της αποτελεσματικότητας μια εκστρατείας για την απόκτηση πελατών και στην ελαχιστοποίηση του κόστους

Αύξηση της αξίας των υφιστάμενων πελατών:

- **Σταυροειδείς πωλήσεις:** Η εξόρυξη δεδομένων, χρησιμοποιώντας τις πληροφορίες των πελατών στις βάσεις δεδομένων, είναι σε θέση να βοηθήσει τον εκπρόσωπο εξυπηρέτησης πελατών να προτείνει τα κατάλληλα επιπρόσθετα προϊόντα στον πελάτη, ή και να μην προτείνει κανένα προϊόν, εάν ο πελάτης δεν είναι δεκτικός σε τέτοιου είδους πωλήσεις. Επίσης, μέσω της εξόρυξης δεδομένων μπορούν να ελαχιστοποιηθούν τα παράπονα των πελατών και να αυξηθεί η κερδοφορία της επιχείρησης.

- **Εξατομίκευση πελατών:** Μέσω της συσταδοποίησης της εξόρυξης δεδομένων γίνεται εφικτή η ομαδοποίηση των παρεμφερών προϊόντων, έτσι ώστε κάθε φορά που κάποιος πελάτης δείχνει ενδιαφέρον για ένα προϊόν, η εταιρεία να προβεί στις συστάσεις για αγορά περισσότερων προϊόντων. Με βάση το προφίλ του πελάτη εντοπίζονται οι πελάτες που μπορεί να ενδιαφέρονται για νέα προϊόντα που προστίθενται στον κατάλογο.

- **Διατήρηση των κερδοφόρων πελατών**: Για σχεδόν κάθε εταιρεία, το κόστος απόκτησης ενός νέου πελάτη υπερβαίνει το κόστος διατήρησης κερδοφόρων πελατών. Με τη δημιουργία του προφίλ των επικερδών, αλλά και των μη επικερδών πελατών, καθίσταται εφικτότερη η διατήρησή τους ως πελάτες, και ο εντοπισμός των πελατών που δεν αποφέρουν σημαντικά έσοδα στην επιχείρηση, αλλά θα μπορούσαν να αποφέρουν στο μέλλον.

Στο παρακάτω σχήμα παρουσιάζουμε τον κύκλο ζωής του πελάτη.

Ο κύκλος ζωής του πελάτη αποτελεί ένα καλό πλαίσιο για την εφαρμογή της εξόρυξης δεδομένων στο CRM, αφού είναι σε θέση να προβλέψει την κερδοφορία των δυνητικών πελατών, καθώς αυτοί γίνονται ενεργοί, πόσο καιρό θα είναι ενεργοί πελάτες και ποιες είναι οι πιθανότητες να παύσουν να είναι πελάτες. Βέβαια, δε θα είναι ένας ακριβής προγνωστικός δείκτης για το πότε συμβαίνουν οι περισσότερες εκδηλώσεις του κύκλου ζωής, αλλά θα μπορεί να βοηθήσει την επιχείρηση να αναγνωρίζει πρότυπα στα δεδομένα των πελατών της, τα οποία είναι προβλέψιμα. Για παράδειγμα, θα μπορούσε να προβλέψει τη συμπεριφορά που περιβάλλει ένα ιδιαίτερο γεγονός του κύκλου ζωής (π.χ. συνταξιοδότηση) και να βρει άλλους πελάτες σε παρόμοια στάδια ζωής και, κατ' επέκταση, ανάλογες συμπεριφορές.

2.7.6 MARKETING DATA INTELLIGENCE

Το αποτέλεσμα της εφαρμογής της εξόρυξης δεδομένων στον κύκλο ζωής του πελάτη είναι το Marketing Data Intelligence, το οποίο ορίζεται ως «ο συνδυασμός του μάρκετινγκ με βάση τα δεδομένα, της τεχνολογίας με σκοπό την μεγιστοποίηση της γνώσης και της κατανόησης των πελατών, και τον συνδυασμό των προϊόντων και δεδομένων συναλλαγής με σκοπό τη βελτίωση της στρατηγικής λήψης αποφάσεων» (Rygielski et al. 2002). Υπάρχουν δύο σημαντικά συστατικά στοιχεία του Marketing Data Intelligence, ο μετασχηματισμός των δεδομένων των πελατών (customer data transformation) και η ανακάλυψη γνώσης των πελατών (customer knowledge discovery).

Τα ακατέργαστα δεδομένα που εξάγονται και μετασχηματίζονται από ένα ευρύ φάσμα εσωτερικών και εξωτερικών βάσεων δεδομένων και η συλλογή όλων αυτών των δεδομένων σε μία κεντρική θέση, όπου μπορούν να αναζητηθούν και να διερευνηθούν, αποτελούν τον μετασχηματισμό των δεδομένων. Η διαδικασία αυτή συνεχίζεται μέσω της ανακάλυψης γνώσης των πελατών, κατά την οποία προέρχονται οι πληροφορίες και μπορούν να εξαχθούν χρήσιμα πρότυπα και συμπεράσματα μέσα από τα δεδομένα. Η διαδικασία αυτή θα πρέπει να παρακολουθείται, έτσι ώστε να διασφαλίζεται ότι τα αποτελέσματα παράγουν αξιοποιήσιμες πληροφορίες (Rygielski et al., 2002).

2.7.7 ΤΜΗΜΑΤΟΠΟΙΗΣΗ ΠΕΛΑΤΩΝ ΜΕ ΕΡΓΑΛΕΙΑ ΤΗΣ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ

Οι παραδοσιακές μέθοδοι τμηματοποίησης πελατών στηρίζονται κυρίως σε εμπειρικές ταξινομήσεις ή απλές στατιστικές μεθόδους. Κατηγοριοποιούν τους πελάτες σύμφωνα με κάποιο κοινό χαρακτηριστικό ή ιδιότητα και δεν μπορούν να κάνουν πιο πολύπλοκες αναλύσεις. Επίσης, καθώς η συσσώρευση των δεδομένων γίνεται ολοένα και μεγαλύτερη, οι παραδοσιακές μέθοδοι δεν μπορούν να αντεπεξέλθουν με το μέγεθος της πολυπλοκότητας (Chen et al., 2006).

Πολλές μεγάλες εταιρείες σήμερα κατέχουν terabytes δεδομένων. Η εξόρυξη δεδομένων δίνει τη δυνατότητα στις επιχειρήσεις να εξάγουν όσο το δυνατόν περισσότερες και αξιοποιήσιμες πληροφορίες μέσα από τον τεράστιο, αυτό, όγκο δεδομένων. Πρόκειται, δηλαδή, για μια λύση σε ένα μεγάλο πρόβλημα που αντιμετωπίζουν πολλές επιχειρήσεις: την υπεραφθονία των δεδομένων και μια σχετική έλλειψη κατάλληλου προσωπικού, τεχνολογίας και χρόνου, ώστε να μετατραπούν τα απλά δεδομένα σε ουσιαστικές και αξιοποιήσιμες πληροφορίες σχετικά με τους υπάρχοντες και μελλοντικούς πελάτες (Chen et al., 2006). Οι τεχνικές εξόρυξης δεδομένων που χρησιμοποιούνται σήμερα στην τμηματοποίηση πελατών ανήκουν στην κατηγορία της συσταδοποίησης (clustering) ή των αλγορίθμων των πλησιέστερων γειτόνων (nearest-neighbors algorithms) (Bounsaythip and Runsala 2001).

Η τμηματοποίηση πελατών στο άμεσο μάρκετινγκ έχει καταστεί ιδιαίτερα αποτελεσματική, λόγω της ανάπτυξης των εφαρμογών μάρκετινγκ στις βάσεις δεδομένων. Οι προσεγγίσεις της εξόρυξης δεδομένων παρέχουν αποδοτικούς τρόπους για τον διαχωρισμό των υφιστάμενων πελατών στα τμήματα αλλά και την ανάπτυξη στρατηγικών μάρκετινγκ, προσαρμοσμένες στα συγκεκριμένα τμήματα ή άτομα. Οι τεχνικές μάρκετινγκ στις βάσεις δεδομένων έχουν εξελιχθεί από τα απλά RFM μοντέλα (Recency, Frequency and Monetary), τα οποία περιλαμβάνουν την πρόσφατη πείρα των αγορών των πελατών, τη συχνότητα των αγορών τους, καθώς και το ποσό των χρημάτων που έχουν δαπανήσει με την επιχείρηση, σε στατιστικές τεχνικές όπως η Chisquare τεχνική ανίχνευσης περιοχών ενδιαφέροντος (chisquare automatic interaction detection - CHAID) και το λογιστικό μοντέλο παλινδρόμησης (logistic regression).

Επίσης, την εμφάνισή τους στις εφαρμογές του μάρκετινγκ στις βάσεις δεδομένων έχουν κάνει και τα νευρωνικά δίκτυα (Yang, 2004). Τα εργαλεία της εξόρυξης δεδομένων παρέχουν ολοκληρωμένη υποστήριξη στις διαδικασίες του management για την απόκτηση και διατήρηση πελατών, την αύξηση της αξίας που προσφέρουν στην επιχείρηση, την πελατειακή ικανοποίηση, καθώς επίσης και την προώθηση της αφοσίωσης των πελατών.

2.7.8 ΜΟΝΤΕΛΟ ΤΜΗΜΑΤΟΠΟΙΗΣΗ ΠΕΛΑΤΩΝ

Η δημιουργία μιας σχεδιασμένης σχέσης ανάμεσα στα αρχικά χαρακτηριστικά του πελάτη αποτελεί το βασικό βήμα για την πελατειακή τμηματοποίηση, κάνοντας χρήση της εξόρυξης δεδομένων. Τα πελατειακά δεδομένα περιέχουν διανεμημένα και συνεχή χαρακτηριστικά. Θέτοντας το κάθε χαρακτηριστικό του πελάτη ως μια διάσταση (dimension) και θέτοντας κάθε πελάτη ως ένα σωματίδιο (particle), όλοι οι πελάτες μαζί μιας επιχείρησης δημιουργούν

έναν πολυδιάστατο χώρο, ο οποίος έχει οριστεί ως «ο χαρακτηριστικός χώρος του κάθε πελάτη» (attribute space of the customer) (Chen et al., 2006).

Η πελατειακή τμηματοποίηση, με χρήση τεχνικών εξόρυξης δεδομένων, είναι δυνατή με τη βοήθεια της λειτουργικής ανάλυσης (functional analysis). Η λειτουργική ανάλυση περιλαμβάνει την ανάλυση της αξίας του πελάτη, την ανάλυση της απόδοσης, την ανάλυση της προώθησης κ.λ.π., βασιζόμενη στα ευρήματα από την σχεδιασμένη σχέση μεταξύ του πελάτη και των βασικών χαρακτηριστικών. Επιπρόσθετα ευρήματα θα ανακαλυφθούν με την ανάπτυξη των πρακτικών management στο CRM. Τα νέα αυτά ευρήματα θα προστεθούν στις θεμελιώδεις διαστάσεις, ενώ η σχεδίαση σχέσεων με τα χαρακτηριστικά των πελατών θα αναδομηθεί.

2.7.8.1 ΠΛΕΟΝΕΚΤΗΜΑΤΑ ΤΗΣ ΜΕΘΟΔΟΥ

Οι παραδοσιακές μέθοδοι τμηματοποίησης πελατών κατατάσσουν τους πελάτες σε κατηγορίες σύμφωνα με τα απλά χαρακτηριστικά των πελατών ή των αγοραζόμενων προϊόντων, όπως για παράδειγμα η κατηγορία του προϊόντος που αγοράστηκε ή η περιοχή κατοικίας. Επίσης, δεν έχουν τη δυνατότητα να εφαρμόζουν σύνθετες αναλύσεις, όπως τον εντοπισμό των πελατών που έχουν υψηλή δυνητική αξία. Τα μοντέλα εξόρυξης δεδομένων πλεονεκτούν έναντι των παραδοσιακών μεθόδων τμηματοποίησης, προσφέροντας πολλά πλεονεκτήματα, τα οποία είναι τα ακόλουθα (Chen et al. 2006):

- **Βελτίωση των προωθητικών ενεργειών:** Το μοντέλο αυτό μπορεί να βοηθήσει την επιχείρηση να ακολουθήσει τις κατάλληλες προωθητικές στρατηγικές, στον κατάλληλο χρόνο και με τα κατάλληλα προϊόντα και υπηρεσίες, με στόχο τους κατάλληλους πελάτες.

- **Ανάλυση της αξίας και της αφοσίωσης των πελατών:** Οι δύο αυτές μεταβλητές είναι πολύ σημαντικές γιατί επηρεάζουν τη στρατηγική της επιχείρησης και τις τακτικές διαχείρισης. Οι επιχειρήσεις μέσω του μοντέλου αυτού μπορούν να διαχωρίσουν σε βαθμίδες τους πελάτες σύμφωνα με την αναμενόμενη αξία τους και την αφοσίωσή τους στην εταιρεία.

- **Ανάλυση του πιστωτικού κινδύνου:** Η αξιολόγηση του κινδύνου είναι ένας αποτελεσματικός τρόπος για την αξιολόγηση ορισμένων συγκεκριμένων τύπων κινδύνου των πελατών, όπως είναι ο κίνδυνος αθέτησης των υποχρεώσεών τους.

- **Δημιουργία νέων προϊόντων έρευνας και ανάπτυξης:** Οι επιχειρήσεις μπορούν να ανακαλύψουν τις προτιμήσεις των πελατών τους μέσω της ανάλυσης πελατών με βάση την εξόρυξη δεδομένων, και να σιγουρευτούν ότι θα υπάρξει συγκεκριμένη ζήτηση για το σχεδιαζόμενο προϊόν.

- **Η επιβεβαίωση της αγοράς στόχου**: Η τμηματοποίηση πελατών με βάση την εξόρυξη δεδομένων είναι σε θέση να εντοπίσει ρητά την αγορά στην οποία απευθύνεται το προϊόν-υπηρεσία της επιχείρησης και να καταστήσει σαφείς τις στοχευμένες πελατειακές ομάδες.

2.7.9 ΔΗΜΙΟΥΡΓΙΑ ΠΡΟΦΙΛ ΠΕΛΑΤΩΝ ΜΕ ΕΡΓΑΛΕΙΑ ΤΗΣ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ

Η δημιουργία προφίλ πελατών είναι ένας τρόπος εφαρμογής εξωτερικών δεδομένων σε ένα πληθυσμό πιθανών πελατών. Ανάλογα με τα διαθέσιμα στοιχεία, αυτά μπορούν να χρησιμοποιηθούν για να προσελκυστούν νέοι πελάτες ή να εγκαταλειφθούν υπάρχοντες «κακοί» πελάτες, αυτοί δηλαδή που δεν πραγματοποιούν επαναλαμβανόμενες αγορές και συνεπώς δεν αποφέρουν σημαντικά έσοδα στην επιχείρηση. Ο στόχος είναι να προβλεφθούν συμπεριφορές με βάση τις πληροφορίες που υπάρχουν για κάθε πελάτη. Μακροπρόθεσμος στόχος της δημιουργίας προφίλ του πελάτη είναι να μετατρέψει την κατανόηση των χαρακτηριστικών της πελατειακής βάσης σε μια αυτοματοποιημένη αλληλεπίδραση με αυτούς (Thearling 2000).

Για το λόγο αυτό, οι ανάγκες της σημερινής αγοράς απαιτούν ένα ευρύ φάσμα διαδικασιών και πληροφοριακών εργαλείων. Αυτά τα εργαλεία χρησιμοποιούνται στη συλλογή δεδομένων και την απλοποίηση της διαδικασίας εξόρυξης γνώσης. Τα εργαλεία εξόρυξης δεδομένων χρησιμοποιούνται για να εντοπίσουν ομάδες με κοινά χαρακτηριστικά στα ιστορικά δεδομένα, όπως για παράδειγμα κριτήρια επιλογής για λίστες αλληλογραφίας ή για να εντοπιστούν αγορές με υψηλές δυνατότητες επέκτασης. Εν συντομία, τα εργαλεία εξόρυξης δεδομένων είναι σε θέση να εντοπίσουν, μέσα από τα δεδομένα, ανθρώπινα ερμηνεύσιμα μοτίβα, όπως επίσης και να χρησιμοποιούν κάποιες μεταβλητές για να προβλέψουν άγνωστες ή μελλοντικές αξίες άλλων μεταβλητών (Bounsaythip and Runsala 2001).

2.7.10 ΑΞΙΟΛΟΓΗΣΗ ΤΗΣ ΑΦΟΣΙΩΣΗΣ ΤΩΝ ΠΕΛΑΤΩΝ

Η δημιουργία μιας πιστής πελατειακής βάσης (customer loyalty) δεν αφορά μόνο στη διατήρηση του αριθμού των πελατών με την πάροδο του χρόνου, αλλά και στην Ενίσχυση της σχέσης της εταιρείας με τους πελάτες, η οποία θα ενθαρρύνει τις μελλοντικές αγορές τους και θα αυξήσει το επίπεδο αφοσίωσης και υποστήριξης. Γνωρίζοντας το επίπεδο αφοσίωσης των πελατών της, η επιχείρηση είναι σε θέση να κατανοήσει το πώς οι προσπάθειές της για τη διατήρηση καλών σχέσεων μπορούν να συμβάλουν στο επίπεδο των κερδών της (Hosseini et al. 2010). Κατά τον Payne (2002), το κόστος για την απόκτηση ενός νέου πελάτη είναι πέντε φορές μεγαλύτερο από το κόστος διατήρησης ενός παλιού πελάτη. Σύμφωνα με τους Eriksson και Vaghult (2000), υπάρχουν πολλές μεταβλητές που επηρεάζουν το επίπεδο πιστότητας των πελατών και κατά συνέπεια, την παραμονή τους ως πελάτες. Οι μεταβλητές αυτές είναι η ποιότητα της σχέσης, η εμπιστοσύνη, το ποσοστό συμμετοχής, η ικανοποίηση, η εξέλιξη της αγοράς, οι οργανωτικές αλλαγές και το κόστος αλλαγής προμηθευτή.

ΚΕΦΑΛΑΙΟ 3
ΕΠΙΧΕΙΡΗΜΑΤΙΚΗ ΕΥΦΥΪΑ ΚΑΙ ΕΞΟΡΥΞΗ ΓΝΩΣΗΣ

3.1 ΟΡΙΣΜΟΣ ΕΠΙΧΕΙΡΗΜΑΤΙΚΗ ΕΥΦΥΪΑΣ

Η έννοια επιχειρηματική ευφυΐα (business intelligence), αναφέρεται σε τεχνολογίες, εφαρμογές, ικανότητες και πρακτικές που χρησιμοποιούνται για να βοηθήσουν επιχειρήσεις στην καλύτερη κατανόηση της αγοραστικής συμπεριφοράς και στην εύρεση επιχειρηματικών ευκαιριών. Κατά τη διάρκεια των προηγούμενων δεκαετιών, οι επιχειρηματικές κινήσεις και αποφάσεις στηρίζονταν σε περιορισμένα δεδομένα και στο επιχειρηματικό ένστικτο του εκάστοτε διευθυντή ή προέδρου. Σήμερα, ο τεράστιος όγκος ηλεκτρονικών δεδομένων, που συλλέγονται πλέον με αυτοματοποιημένο τρόπο, σε συνδυασμό με τις τεχνολογικές εξελίξεις, έχει προκαλέσει την ανάγκη για ειδικά εργαλεία διαχείρισης και εκμετάλλευσης της επιχειρηματικής πληροφορίας.

3.2 ΕΡΓΑΛΕΙΑ ΕΠΙΧΕΙΡΗΜΑΤΙΚΗΣ ΕΥΦΥΪΑΣ

Ως εργαλεία επιχειρηματικής ευφυΐας ορίζονται συγκεκριμένες εφαρμογές λογισμικού (application software) σχεδιασμένες να αναλύουν και να απεικονίζουν δεδομένα, αλλά και να παράγουν χρήσιμες αναφορές (reports). Συνδέονται συνήθως άμεσα με τη βάση ή αποθήκη ηλεκτρονικών δεδομένων μιας εταιρίας.

Τα εργαλεία επιχειρηματικής ευφυΐας χωρίζονται στις παρακάτω βασικές κατηγορίες:

- ✓ Spreadsheets
- ✓ Προγράμματα αναφορών και ερωτημάτων, λογισμικό για εξαγωγή, ταξινόμηση και παρουσίαση δεδομένων επιλεκτικά
- ✓ Κύβοι OLAP
- ✓ Digital Dashboards
- ✓ Data mining
- ✓ Predictive analytics
- ✓ Business performance management

Αυτά τα εργαλεία μπορεί να τα προμηθευτεί κανείς είτε αυτόνομα, είτε σαν σουίτες, είτε σαν επιπρόσθετα τμήματα σε υπάρχοντα προγράμματα διαχείρισης (π.χ. ERP) ή σε βάσεις δεδομένων.

ΚΕΦΑΛΑΙΟ 4
ΣΥΝΔΕΣΜΟΙ ΠΟΥ ΑΦΟΡΟΥΝ ΤΗΝ ΕΞΟΡΥΞΗ ΓΝΩΣΗΣ

4.1 ΕΡΓΑΛΕΙΑ ΕΛΕΥΘΕΡΟΥ ΛΟΓΙΣΜΙΚΟΥ (OPEN SOURCE)

AlphaMiner: Το AlphaMiner είναι μια πλατφόρμα βασισμένη σε Java, για εφαρμογές εξόρυξης δεδομένων που αναπτύχθηκε στο πανεπιστήμιο του Hong Kong.

RapidMiner: Εφαρμογή βασισμένη σε ελεύθερο λογισμικό για ανάλυση γνώσης και δεδομένων. Μπορεί να ενσωματωθούν σε άλλες εφαρμογές ή προϊόντα.

Weka – Machine Learning Techniques: Το Weka είναι μια συλλογή εργαλείων και τεχνικών μάθησης για εφαρμογές εξόρυξης γνώσης, και όχι μόνο, το οποίο έχει αναπτυχθεί σε γλώσσα Java και ο κώδικας είναι ανοικτός στο κοινό.

MLcomp: Το MLcomp είναι ένας διαδραστικός ιστοχώρος αντικειμενικής σύγκρισης αλγορίθμων πάνω σε διαφορετικά σύνολα δεδομένων.

Hadoop: Το Apache Hadoop είναι ένα ανοιχτού κώδικα λογισμικό (framework) που υποστηρίζει τη παράλληλη επεξεργασία μεγάλου όγκου δεδομένων, διανέμοντας τμήματα της εφαρμογής σε πολλές συστάδες (clusters) υπολογιστών.

4.2 ΕΤΑΙΡΕΙΕΣ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ

Oracle: Δημοφιλής και αξιόπιστη βάση δεδομένων η οποία υποστηρίζει μεταξύ άλλων, λειτουργίες εξόρυξης γνώσης και ανάλυσης δεδομένων.

SPSS: Επιχειρηματικές λύσεις και παροχή λογισμικού για ανάλυση δεδομένων.

Microsoft: Οι νεότερες εκδόσεις του SQL (Structured Query Language) περιλαμβάνουν μια ειδική πλατφόρμα για business intelligence με δυνατότητες εξόρυξης δεδομένων, οι οποίες επιτρέπουν σε οργανισμούς και εταιρείες να πραγματοποιούν ανάλυση δεδομένων και να αξιοποιούν τα αποτελέσματα μέσα από το γνώριμο περιβάλλον του Microsoft Office.

IBM: Η IBM προσφέρει εργαλεία επιχειρηματικής ευφυΐας και εξόρυξης γνώσης για υψηλών απαιτήσεων αναλύσεις και προβλέψεις.

TARGIT: Η TARGIT είναι μια εταιρεία με 25 και πλέον χρόνια εμπειρίας στις τεχνολογίες αιχμής στο χώρο της επιχειρηματικής ευφυΐας. Η TARGIT Greece αποτελεί τον αντιπρόσωπο του BI software για την Ελλάδα.

ΚΕΦΑΛΑΙΟ 5
BIG DATA ΚΑΙ ΕΞΟΡΥΞΗ ΓΝΩΣΗΣ

5.1 ΤΑ «ΜΕΓΑΛΑ ΔΕΔΟΜΕΝΑ» ΑΛΛΑΖΟΥΝ ΤΑ ΔΕΔΟΜΕΝΑ.

Δεν υπάρχει αμφιβολία ότι το 2012 ήταν η χρονιά των μεγάλων δεδομένων, αρκεί να πληκτρολογήσει κανείς στη μηχανή αναζήτησης Google τις λέξεις big data, για να αντιληφθεί πόσο μας άλλαξαν, αλλά και πόσο θα μας αλλάξουν τη ζωή. Σύμφωνα με έρευνα του Gartner, αυτή είναι μόνο η αρχή. Το 2012 ξοδεύτηκαν 96 δισεκατομμύρια δολάρια παγκοσμίως σε επενδύσεις για «big data», ποσό που αναμένεται να αγγίξει τα 232 δισεκατομμύρια το 2016.

Ο ορισμός της έννοιας μεγάλα δεδομένα σύμφωνα με το Gartner το 2012 είναι: «Τα big data είναι υψηλού όγκου, υψηλής ταχύτητας ή υψηλής ποικιλίας στοιχεία που απαιτούν αποδοτικές και καινοτόμες μορφές επεξεργασίας πληροφοριών».

Στα μεγάλα δεδομένα συγκαταλέγονται όλες οι πληροφορίες των social media που είναι προσβάσιμες σε όλους μας και βρίσκονται στο διαδίκτυο, δηλαδή φωτογραφίες, video και κείμενα, καθώς και όλα τα κλειστά δεδομένα των διαφόρων εταιρειών αλλά και των κυβερνήσεων.

Big data, ο μελλοντικός πυρετός του χρυσού.

Τα big data μπορούν να αλλάξουν τον τρόπο που επικοινωνούν οι επιχειρήσεις μεταξύ τους αλλά και τον τρόπο που επικοινωνούμε όλοι μεταξύ μας. Ο δημοσιογράφος και επιχειρηματίας Γκρεγκ Χάντφιλντ που διοργάνωσε πριν κάποιους μήνες το συνέδριο Open-Data Cities, στο Μπράιτον της Μεγάλης Βρετανίας δηλώνει ότι τα μεγάλα δεδομένα πρέπει να είναι διαθέσιμα σε όλους. Το Open-Data Cities προσπαθεί να εκδημοκρατίσει δεδομένα του δημόσιου τομέα, ώστε να προωθηθούν καινοτομίες στον ιδιωτικό τομέα. Η προσπάθεια αυτή αφορά χωριά, πόλεις και χώρες όλου του κόσμου.

Στο συνέδριο Open-Data Cities ο Μπιλ Τόμσον του BBC παραλλήλισε την προσπάθεια του Open-Data Cites με την κατασκευή της Βενετίας, λέγοντας χαρακτηριστικά: «Κάποια στιγμή θα φτιάξουμε μεγαλοπρεπή παλάτια, αλλά προς το παρόν σφυρηλατούμε κούτσουρα σε ένα βάλτο». Αν και πολλοί μπορεί να θεωρούν τα big data τον μελλοντικό πυρετό του χρυσού, ίσως είναι καλύτερα να κοιτάξουμε το παρόν και να παραδειγματιστούμε από κάποιες εταιρίες που τα αξιοποιούν ήδη με μεγάλη δημιουργικότητα.

Για παράδειγμα, η εταιρία τηλεφωνίας IOVOX που εδρεύει στο Λονδίνο, έφτιαξε μια πλατφόρμα που δίνει πληροφορίες σε πραγματικό χρόνο για την κίνηση των τηλεφωνημάτων. Τα μεγάλα δεδομένα είναι ένα από τα πιο υποτιμημένα εργαλεία για τις μοντέρνες επιχειρήσεις. Ο κόσμος ξεχειλίζει από πληροφορίες, αλλά το πρόβλημα που αντιμετωπίζουμε είναι ότι δεν έχουμε τη δυνατότητα να ξεχωρίζουμε τα χρήσιμα δεδομένα από το θόρυβο, δήλωσε ο Ράϊαν Γκάλαχερ, Διευθύνων Σύμβουλος και ιδρυτής της IOVOX.

Η πρώτη διδάξασα εταιρία που αξιοποίησε τα μεγάλα δεδομένα και το διαδίκτυο, ήταν η Google, όταν ακόμη κανείς δεν μπορούσε να προβλέψει την επιτυχία που θα είχαν. Τώρα

όμως μπορούμε μετά βεβαιότητας να προβλέψουμε ότι οι επόμενες εταιρίες που θα αξιοποιήσουν αυτή τη δύναμη των μεγάλων δεδομένων και του διαδικτύου θα είναι σίγουρα κερδισμένες.

5.2 ΑΠΟΤΥΓΧΑΝΟΥΜΕ ΝΑ ΑΝΤΛΗΣΟΥΜΕ ΓΝΩΣΗ ΑΠΟ ΤΑ BIG DATA, ΣΥΝΕΧΙΖΟΥΜΕ ΤΗΝ ΕΞΟΡΥΞΗ.

Ένας στους τρεις οργανισμούς που επιχειρούν να εξάγουν ειδικές γνώσεις από την ανάλυση όλων των δεδομένων τους αποτυγχάνουν, δείχνει νέα έρευνα της Hewlett Packard. Η ανάλυση των επονομαζόμενων big data αποδεικνύεται δύσκολη υπόθεση, εντούτοις έξι στους δέκα οργανισμούς δηλώνουν πρόθυμοι να δεσμεύσουν το 10% του προϋπολογισμού τους για την καινοτομία για να εξορύξουν πολύτιμη πληροφορία από δομημένα και αδόμητα δεδομένα.

Ότι κι αν σημαίνει big data, η νεότερη έρευνα της Hewlett Packard τον Μάιο του 2013 επαναλαμβάνει ότι η ανάλυση της διαρκούς ροής δεδομένων δομημένων και αδόμητων παραμένει πρόκληση για τις επιχειρήσεις. Ο αυξανόμενος όγκος, η ποικιλομορφία αλλά και η ευπάθεια των δεδομένων που ρέουν από το εσωτερικό αλλά και το περιβάλλον κάθε επιχείρησης αποτελούν την αιτία της αποτυχίας των πρωτοβουλιών big data. Περισσότεροι από έναν στους τρεις οργανισμούς που έχει επιχειρήσει την ανάλυση έχει αποτύχει, δείχνει η έρευνα που πραγματοποιήθηκε για λογαριασμό της Hewlett Packard υπό τον τίτλο "Big Data and Cloud" από την Coleman Parkers Research Ltd., τον Μάϊο του 2013.

Σε προγενέστερη παγκόσμια έρευνα που εκδόθηκε τον Απρίλιο και διενεργήθηκε επίσης για λογαριασμό της HP, περισσότερα από ένα στα δύο στελέχη επιχειρήσεων ανέφεραν ότι οι οργανισμοί τους δεν είναι εξοπλισμένοι με τις σωστές λύσεις για να αντλήσουν ειδικές γνώσεις από τα Big Data. Επιπλέον, δεν διαθέτουν την τεχνογνωσία καθώς και τη συνεκτική στρατηγική για να συγκεντρώσουν όλα τα στοιχεία και στη συνέχεια να ενσωματώσουν νέα αλλά και παλιά δεδομένα. Παρά τις συνεχείς αποτυχίες, το 60% των εταιρειών που συμμετείχαν στην νεότερη έρευνα δηλώνουν πρόθυμες να δεσμεύσουν για το big data το 10% του προϋπολογισμού τους για την καινοτομία. Αυτό δείχνει ότι έχουν πειστεί ότι τα επονομαζόμενα big data κρύβουν πλούσιες γνώσεις και η ανάλυσή τους μπορεί να προσφέρει αποτελέσματα σε πραγματικό χρόνο. Με την εξαγωγή ειδικών γνώσεων που είναι κρυμμένες μέσα στα big data οι επιχειρήσεις μπορούν να εξορθολογήσουν τις βασικές οργανωτικές διαδικασίες όπως, οι προσφορές, οι προμήθειες, η εφοδιαστική αλυσίδα και οι λειτουργίες απογραφής, υποστηρίζουν εταιρείες όπως η Hewlett Packard, η IBM και άλλες εταιρείες που παρέχουν hardware, software και υπηρεσίες.

«Τα big data επιτρέπουν στους οργανισμούς να επωφεληθούν από το σύνολο των πληροφοριών τους, τόσο των εσωτερικών όσο και των εξωτερικών, σε πραγματικό χρόνο. Παράγουν εξαιρετικά γρήγορη διαδικασία λήψης αποφάσεων και ως αποτέλεσμα, μοναδικούς και καινοτόμους τρόπους για την προσφορά υπηρεσιών στους πελάτες και την κοινωνία», δήλωσε ο George Kadifa, επικεφαλής του τμήματος software στην HP κατά την ανακοίνωση του HAVEn. Το HAVEn αποτελεί την νεότερη πρόταση της HP για την ανάλυση big data, μια πλατφόρμα big data analytics, η οποία αξιοποιεί το λογισμικό, το hardware και τις υπηρεσίες analytics της HP (συνδυάζει τις δοκιμασμένες τεχνολογίες των HP Autonomy, HP Vertica, HP ArcSight και HP Operations Management). Η Hewlett Packard υποστηρίζει πως με αυτή την πλατφόρμα οι επιχειρήσεις μπορούν να αντλήσουν αξία

από το 100% των πληροφοριών, συμπεριλαμβανομένων των δομημένων, των ημι-δομημένων και των μη-δομημένων δεδομένων.

Περνώντας από τις υψηλού επιπέδου έννοιες των big data στην πραγματικότητα, δηλαδή στην αξιοποίησή τους και τη δράση με βάση την πληροφορία, οι ειδικές γνώσεις που προκύπτουν από την ανάλυση τους μπορούν να βοηθήσουν μια επιχείρηση να βελτιώσει:

- Το cross-channel marketing σε πραγματικό χρόνο
- Την εμπειρία του πελάτη και την πρόβλεψη της ζήτησης

Άλλες περιπτώσεις χρήσης περιλαμβάνουν:

- Την υποστήριξη των προσπαθειών των πελατών να εντοπίσουν και να προλάβουν την απάτη,
- Να διασφαλίσουν τη συμμόρφωση σε πραγματικό χρόνο
- Να χρησιμοποιήσουν τα κοινωνικά δίκτυα για να διαχειριστούν τους κινδύνους και τη φήμη της μάρκας τους.

Είναι γεγονός ότι ο όρος big data έχει πολλούς ορισμούς, όπως άλλωστε και η πρακτική αξιοποίησή της ανάλυσης των πάντοτε αδιευκρίνιστων δεδομένων.

ΣΥΜΠΕΡΑΣΜΑΤΑ

Στη παρούσα εργασία ασχοληθήκαμε με την επιστήμη της εξόρυξης γνώσης και των δεδομένων, κάναμε μια εισαγωγή στις βασικές της έννοιες, αναφερθήκαμε στους στόχους, τα στάδια, τις κατηγορίες και τις μεθόδους της επιστήμης αυτής, ενώ απαραίτητη ήταν και η αναφορά στα πλεονεκτήματα και τα μειονεκτήματα που απορρέουν από τη συγκεκριμένη επιστήμη. Σημαντική ήταν επίσης και η αναφορά στα εργαλεία και τα μοντέλα που χρησιμοποιεί η εξόρυξη γνώσης, την αποτελεσματικότητα και χρησιμότητα τους και το κατά πόσο αυτά μπορούν να υλοποιηθούν από τις όλο και πιο ανταγωνιστικές επιχειρήσεις της σύγχρονης εποχής.

Καταλήγοντας λοιπόν, μπορούμε να πούμε ότι η εξόρυξη γνώσης από βάσεις δεδομένων, αποτελεί μία από τις πλέον σύγχρονες τεχνικές εξόρυξης πληροφορίας από ακατέργαστα δεδομένα. Πιο αναλυτικά, ψάχνει δομές και πληροφορίες σε ένα τεράστιο όγκο δεδομένων με σκοπό την μετατροπή αυτών σε χρήσιμη πληροφορία. Οι εφαρμογές της εξόρυξης δεδομένων εκτείνονται σε όλο το φάσμα των επιστημών από Πληροφορική, Οικονομία, Εκπαίδευση μέχρι Βιολογία και Αστρονομία.

Όλες τις πληροφορίες για την παρούσα εργασία τις αντλήσαμε από ελληνική και ξένη βιβλιογραφία, ενώ σημαντική ήταν και η συνεισφορά πηγών μέσω διαδικτύου. Οι συνεχείς έρευνες και οι όλο και πιο εξονυχιστικές μελέτες πάνω στην επιστήμη της εξόρυξης γνώσης, έχουν σαν αποτέλεσμα να την καθιστούν σαν ένα από τα πλέον απαραίτητα εργαλεία στη σημερινή κοινωνία.

ΒΙΒΛΙΟΓΡΑΦΙΑ

Α. ΕΛΛΗΝΙΚΗ

Θεοδωρίδης Γ., Πελέκης Ν. (2011): Εξόρυξη Γνώσης από Δεδομένα - Συσταδοποίηση, Ομάδα Διαχείρισης Δεδομένων Πανεπιστήμιο Πειραιώς

Σαλατάς Ι. (2011): Υλοποίηση και εφαρμογή Τεχνητών Νευρωνικών Δικτύων για την πρόβλεψη χρονοσειρών συναλλαγματικών ισοτιμιών, Ελληνικό Ανοιχτό Πανεπιστήμιο.

Σταυλιώτης Ε. Γεράσιμος (2009): Εξόρυξη Δεδομένων και Αναγνώριση προτύπων σε κατηγορικά δεδομένα μέσω συσταδοποίησης, Ελληνικό Στατιστικό Ινστιτούτο

Κωνσταντίνος Δ. (2007): Τεχνητά Νευρωνικά Δίκτυα., Εκδόσεις Κλειδάριθμος, Αθήνα

Β. ΞΕΝΟΓΛΩΣΣΗ

Hall A. M., Frank E., Witten H. I (2011): Data Mining, Practical Machine Learning Tools and Techniques

Lazarevic A., (2008): Data Mining for Anomaly Detection, Tutorial at the European

Kumar V., Steinbach M. (2006): Introduction to Data Mining, Addison Wesley.

Witten I.H. and Frank E., (2005): Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann

Kumar, Steinbach, Tan (2004): Introduction to Data Mining, University of Stanford

Dunham M.H., (2004): Data Mining introductory and advanced topics", Prentice Hall

Ester M., Sander J. (2000): Knowledge Discovery in Databases. Techniken und Anwendungen, Springer, Berlin.

Prodromidis A. and Chan P., (2000): Meta-learning in distributed data mining systems, Issues and Approaches, in Advances of Distributed Data Mining, AAAI Press.

Murthy S. (1998): Automatic construction of decision trees from data, A multidisciplinary survey. Data Mining and Knowledge Discovery

Freitas, A.A., (1998): A Survey of Parallel Data Mining, in Proceedings of 2nd International Conference on the Practical Applications of Knowledge Discovery and Data Mining Friedman J. H. K., (1997): On bias, variance, 0/1-loss, and the curse-of-dimensionality, Data Mining and Knowledge Discovery

Fayyad U., Piatetsky-Shaprio G., Smyth P. and Uthurusamy R., (1996): Advances in Knowledge Discovery and Data Mining, MIT Press, Cambridge

Lawrence J. (1994): Introduction to Neural Networks, California Scientific Software Press

Ripley, B. D. (1994): Neural Networks and Related Methods for Classification, Journal of the Royal Statistics Society

Γ. ΔΙΑΔΙΚΤΥΟ

http://www.reporter.gr/

http://videolectures.net/

http://www.datamining.gr/

http://www.gartner.com/technology/home.jsp

http://www.uea.ac.uk/computing/data-mining

http://www-users.cs.umn.edu/~aleks/pkdd08.pdf

http://www.stat.tamu.edu/~eparzen/ftp/future.pdf

http://www.euretirio.com/2010/05/egr-euretirio.html

http://www.britannica.com/EBchecked/topic/1056150/data-mining

http://www-stat.stanford.edu/~jtaylo/courses/stats202/notes/chap4_basic_classification.pdf

http://www.britannica.com/EBchecked/media/55236/A-section-of-an-artificial-neural-network-In-the-figure